

Bayesian Repulsive Gaussian Mixture Model

Fangzheng Xie*

Yanxun Xu^{*†}

Abstract

We develop a general class of Bayesian repulsive Gaussian mixture models that encourage well-separated clusters, aiming at reducing potentially redundant components produced by independent priors for locations (such as the Dirichlet process). The asymptotic results for the posterior distribution of the proposed models are derived, including posterior consistency and posterior contraction rate in the context of nonparametric density estimation. More importantly, we show that, as a measurement of the model complexity, the posterior number of necessary components to fit the data well grows sub-linearly with respect to the sample size asymptotically. In addition, an efficient and easy-to-implement blocked-collapsed Gibbs sampler is developed based on the exchangeable partition distribution and the corresponding urn model. We evaluate the performance and demonstrate the advantages of the proposed model through extensive simulation studies and real data analysis. The R code is available at https://drive.google.com/open?id=0B_zFse0eqxBHVMduVEM2Tk9tZFU.

Key Words: Blocked-Collapsed Gibbs Sampler, Model Complexity, Nonparametric Density Estimation, Posterior Convergence, Urn-Model

*Department of Applied Mathematics and Statistics, Johns Hopkins University

[†]Correspondence should be addressed to Yanxun Xu (yanxun.xu@jhu.edu)

1 Introduction

In Bayesian analysis of mixture models, the independent priors on the component-specific parameters have been widely used because of their flexibility and technical convenience. A nonparametric example is the renowned Dirichlet process (DP) where the atoms in the stick-breaking representation are independent and identically distributed (i.i.d.) from a base distribution. One of the potential but non-negligible issues for such an approach is the presence of redundant components, especially when parsimony on the number of components is preferred. For example, when a mixture model is used in biomedical applications, each component of the mixture may be interpreted as clinically or biologically meaningful subpopulations (of patients, disease types, etc.). To address this challenge, in this paper we argue for a Bayesian approach for modeling repulsive mixtures as a competitive alternative, establish its posterior consistency and posterior contraction rate, and study the asymptotic behavior of the posterior number of components.

The mixture models have been extensively studied from both the frequentist and the Bayesian perspectives. Formally, given the parameter space Θ , a mixture model with a kernel density $\psi : \mathbb{R}^p \times \Theta \rightarrow \mathbb{R}_+$ and a mixing distribution $G \in \mathcal{M}(\Theta)$ can be represented as $\mathbf{y}_i \sim \int_{\Theta} \psi(\mathbf{y}, \boldsymbol{\theta}) dG(\boldsymbol{\theta})$, where $\mathcal{M}(\Theta)$ is a class of probability distributions on Θ (equipped with an implicitly specified suitable σ -field). The most commonly used kernel density ψ is the normal density, which leads to the Gaussian mixture model (GMM). In particular, the GMM with a discrete (potentially infinitely supported) mixing $G = \sum_k w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$ has been widely used for clustering, since an equivalent characterization is $\mathbf{y}_i \mid z_i \sim N(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$, $\mathbb{P}(z_i = k) = w_k$, where z_i encodes the clustering membership of the corresponding observation \mathbf{y}_i . The parameters for each component $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 1, \dots, K$, are referred to as the cluster/component-specific parameters. Throughout we use K to denote the (potentially infinite) number of components in a mixture model. Frequentists' ways of modeling mixture

models require a finite and fixed K , the estimation of which could be accomplished using model selection approaches. Nonparametric Bayesian priors allow us to perform inference without *a priori* fixed and finite K . For example, the DP prior on G yields an exchangeable partition distribution on $(\boldsymbol{\theta}_{z_1}, \dots, \boldsymbol{\theta}_{z_n})$, the inference of which indicates a distribution of the number of clusters among $(\boldsymbol{\theta}_{z_1}, \dots, \boldsymbol{\theta}_{z_n})$. The development of Markov chain Monte Carlo sampling techniques (Ishwaran and James, 2001, 2002; Antoniak, 1974; MacEachern and Mueller, 1998; Neal, 2000; Walker, 2007) further popularizes the DP mixture model in a wide array of applications, such as biomedicine, machine learning, pattern recognition, etc.

Meanwhile, the asymptotic results of the DP mixture of Gaussians as a method of nonparametric density estimation have been studied. In the univariate case, the posterior consistency of the DP mixture of univariate Gaussians was established by Ghosal et al. (1999), and the posterior convergence rate was studied by Ghosal and Van Der Vaart (2001). Posterior consistency in the multivariate setting (Wu and Ghosal, 2010) is harder due to the exponential growth of the L_1 -entropy of sieves. Shen et al. (2013); Canale et al. (2017) derived the posterior contraction rates at general smooth densities for multivariate density estimation using the DP mixture of Gaussians.

Nevertheless, as shown in Xu et al. (2016), the DP mixture model typically produces relative large number of clusters, some of which are typically redundant. Theoretically, Miller and Harrison (2013) showed that when the underlying data generating density is a finite mixture of Gaussians, the posterior number of clusters under the DP mixture model is not consistent. In other words, the posterior distribution of the number of clusters does not converge to the point mass at the underlying true K . Alternatively, finite mixture model with a prior on K , referred to as the mixture of finite mixtures(MFM) (Nobile, 1994; Miller and Harrison, 2016), was developed. The posterior inference of the MFM can be carried out either by the reversible-jump Markov chain Monte Carlo (RJ-MCMC) (Green, 1995), or by the collapsed Gibbs sampler derived via the exchangeable partition representation (Miller and

Harrison, 2016). Meanwhile, the posterior asymptotics for the MFM as a density estimator, to our best knowledge, is restricted to the cases of univariate location-scale mixtures (Kruijer et al., 2010) and multivariate location mixtures (Shen et al., 2013), in which the priors on locations are assumed to be conditionally i.i.d. given K .

These approaches, however, assume independent prior on the component-specific parameters $(\theta_1, \dots, \theta_K)$. In the context of parametric inference, where the underlying data generating distribution is a finite mixture of Gaussians, repulsive priors (Petrulia et al., 2012; Quinlan et al., 2017) and non-local priors (Fuquene et al., 2016) were developed as shrinkage methods to penalize mixture models with redundant components. In particular, theoretical properties regarding only univariate parametric density estimations were discussed in Petrulia et al. (2012) and Quinlan et al. (2017). In addition, Xu et al. (2016) proposed repulsive mixtures via determinantal point process (DPP) with a prior on K , where the RJ-MCMC sampler for the posterior inference is potentially inefficient in high-dimensional setting.

In this paper, we propose a Bayesian repulsive Gaussian mixture (RGM) model. The main contributions of this paper are as follows. First, under certain mild regularity conditions, we establish the posterior consistency for nonparametric density estimation under the RGM model, and obtain an “almost” parametric posterior contraction rate $(\log n)^t/\sqrt{n}$ for $t > p + 1$. It is worth mentioning that our theoretical results is a nonparametric extension to those in Petrulia et al. (2012); Quinlan et al. (2017), as we do not assume a parametric form of the underlying data generating density f_0 . Second, the model complexity in terms of the asymptotic relationship between the posterior number of components and the sample size is studied as well. It turns out that in order to fit the data well, it is sufficient that K grows sub-linearly with respect to the sample size. Furthermore, instead of fixing K or implementing a RJ-MCMC sampler for the posterior inference of the RGM model, we develop a more efficient blocked-collapsed Gibbs sampler that is based on the exchangeable

partition distributions.

The remainder of the paper is organized as follows. In Section 2 we formulate the Bayesian repulsive Gaussian mixture model. Section 3 elaborates the asymptotic results for the posterior distribution. In particular, we establish the posterior consistency, investigate posterior contraction rate, and study the asymptotic behavior of the posterior number of components. In section 4 we develop the generalized urn model for the RGM model by integrating out the mixing weights and K , and design an efficient blocked-collapsed Gibbs sampler. Section 5 demonstrates the advantages of the proposed model as well as the efficiency of the proposed inference algorithm via simulation studies and real data analysis. We conclude the paper in Section 6.

2 Bayesian Repulsive Mixture Model

In this section we formulate the RGM model in a Bayesian framework. Suppose $\mathcal{S} \subset \mathbb{R}^{p \times p}$ is a collection of positive definite matrices, equipped with the Borel σ -field on \mathcal{S} . We consider the Gaussian mixture model, a family of densities of the form

$$f_F(\mathbf{y}) = \int_{\mathbb{R}^p \times \mathcal{S}} \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) dF(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (2.1)$$

where $\phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp \left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \right]$ is the density of the p -dimensional Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and F is a distribution on $\mathbb{R}^p \times \mathcal{S}$. We shall also use the shorthand notation $\phi_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}) = \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $f_F = \phi_{\boldsymbol{\Sigma}} * F$, where $*$ is the conventional notation for convolution of two functions. We assume that the data $(\mathbf{y}_n)_{n=1}^\infty$ are i.i.d. generated from some unknown density f_0 , the estimation of which is of interest.

Denote the space of all probability distributions over $\mathbb{R}^p \times \mathcal{S}$ by $\mathcal{M}(\mathbb{R}^p \times \mathcal{S})$, and that over \mathbb{R}^p by $\mathcal{M}(\mathbb{R}^p)$. We define a prior Π on f over the space of all density functions in \mathbb{R}^p

by the following hierarchical model:

$$\begin{aligned}
(f(\mathbf{y}) \mid F) &= \int_{\mathbb{R}^p \times \mathcal{S}} \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) dF(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\
(F \mid K, \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K) &= \sum_{k=1}^K w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}, \\
(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K \mid K) &\sim p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K \mid K), \\
(w_1, \dots, w_K \mid K) &\sim \mathcal{D}_K(\beta), \quad K \sim p_K(K), \quad K \in \mathbb{N}_+.
\end{aligned} \tag{2.2}$$

Here $p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K \mid K) > 0$ is some density function with respect to the Lebesgue measure on $(\mathbb{R}^p \times \mathcal{S})^K$, $\mathcal{D}_K(\beta)$ is the symmetric Dirichlet distribution over Δ^K with density function $p(w_1, \dots, w_K) = \Gamma(K\beta)/\Gamma(\beta)^K \prod_{k=1}^K w_k^{\beta-1}$, where $\Delta^K = \{(w_1, \dots, w_K)^T : \sum_{k=1}^K w_k = 1, w_k \geq 0\}$ is the ℓ_1 -simplex on \mathbb{R}^K . The prior on K that is supported on all positive integers is essential, as we allow the number of components to grow with the sample size in order to fit the data well.

Instead of assuming $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{k=1}^K$ being i.i.d. from a “base measure”, we introduce repulsion among components $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ through their centers $\boldsymbol{\mu}_k$, such that they are well separated. We assume the density $p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K \mid K)$ is of the following form,

$$p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K \mid K) = \frac{1}{Z_K} \left[\prod_{k=1}^K p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_k) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}_k) \right] h_K(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K), \tag{2.3}$$

where $Z_K = \int \dots \int_{\mathbb{R}^p \times \mathcal{S}} h_K(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K) \left[\prod_{k=1}^K p(\boldsymbol{\mu}_k) \right] d\boldsymbol{\mu}_1 \dots d\boldsymbol{\mu}_K$ is the normalizing constant, and the function $h_K : (\mathbb{R}^p)^K \rightarrow [0, 1]$ is invariant under permutation of its arguments: $h_K(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K) = h_K(\boldsymbol{\mu}_{\mathfrak{T}(1)}, \dots, \boldsymbol{\mu}_{\mathfrak{T}(K)})$ for any permutation $\mathfrak{T} : \{1, \dots, K\} \rightarrow \{1, \dots, K\}$. We require that h_K satisfies the following repulsive condition: $h_K(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K) = 0$ if and only if $\boldsymbol{\mu}_k = \boldsymbol{\mu}_{k'}$ for some $k \neq k'$, $k, k' \in \{1, \dots, K\}$. In this paper, we focus on the case where the repulsive property is introduced only through the mean vectors $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)$, *i.e.* we allow nonvanishing density even when distinct components share an identical covariance matrix. The case where repulsion is introduced through the covariance matrices is of

independent interest and may be further explored.

We consider the following two classes of repulsive functions $h_K(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)$:

$$h_K(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K) = \min_{1 \leq k < k' \leq K} g(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|), \quad (2.4)$$

$$h_K(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K) = \left[\prod_{1 \leq k < k' \leq K} g(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|) \right]^{\frac{1}{K}}, \quad (2.5)$$

for $K \geq 2$, and $h_1(\boldsymbol{\mu}_1) \equiv 1$, where $g : \mathbb{R}_+ \rightarrow [0, 1]$ is a strictly monotonically increasing function with $g(0) = 0$. Notice that the repulsive functions defined here generalize those in [Petràlia et al. \(2012\)](#); [Quinlan et al. \(2017\)](#), who fix K due to the challenges in estimating K caused by the complicated relation between Z_K and K . However, for the two repulsive functions (2.4) and (2.5), we are able to find the connection between Z_K and K in **Theorem 1**, the proof of which is deferred to Section B of the Supplementary Material. We will discuss the asymptotic behavior of the posterior distribution of K in Section 3.4.

Theorem 1. *Suppose the repulsive function h_K is either of the form (2.4) or (2.5). If $\iint_{\mathbb{R}^p \times \mathbb{R}^p} [\log g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)]^2 p(\boldsymbol{\mu}_1) p(\boldsymbol{\mu}_2) d\boldsymbol{\mu}_1 d\boldsymbol{\mu}_2 < \infty$, then $0 \leq -\log Z_K \leq c_1 K$ for some constant $c_1 > 0$.*

We refer to the prior Π on $f \in \mathcal{M}(\mathbb{R}^p)$ given by (2.2), (2.3), (2.4) or (2.5) as the Bayesian *repulsive Gaussian mixture* (RGM) model, denoted by $f \sim \text{RGM}_1(\beta; g, p_{\boldsymbol{\mu}}, p_{\boldsymbol{\Sigma}}, p_K)$ if h_K is of the form (2.4), or $f \sim \text{RGM}_2(\beta; g, p_{\boldsymbol{\mu}}, p_{\boldsymbol{\Sigma}}, p_K)$ if h_K is of the form (2.5).

3 Asymptotic Results for Posterior Distribution

In this section we discuss the asymptotic results for the posterior of the RGM model defined in Section 2 in terms of nonparametric density estimation. In particular, we establish the posterior consistency, discuss the posterior contraction rate, and study the posterior number of components in terms of the sample size n . We defer the proofs of all theorems, lemmas,

propositions, and corollaries to Sections [C](#), [D](#), and [E](#) of the Supplementary Material.

3.1 Preliminaries and Notations

We begin with some useful notations. Given a positive definite matrix Σ , we use $\lambda(\Sigma)$ to denote an eigenvalue of Σ , and $\lambda_{\max}(\Sigma)$, $\lambda_{\min}(\Sigma)$ to denote the largest and smallest eigenvalue of Σ , respectively. Denote \mathbf{I} the identity matrix, and $\mathbf{I}_p \in \mathbb{R}^{p \times p}$ the identity matrix of size $p \times p$ if specifying matrix dimension is needed. The Kullback-Leibler (KL) divergence between two densities f and g is denoted by $D(f \parallel g) = \int f \log(f/g)$. Denote $\|\cdot\|$ the Euclidean norm on \mathbb{R}^p . We use $\|\cdot\|_1$ to denote both the L_1 -norm on $L^1(\mathbb{R}^p)$ and the ℓ_1 -norm on finite dimensional Euclidean space \mathbb{R}^d for any $d \geq 1$. $\|\cdot\|_\infty$ is used to denote both the ℓ_∞ -norm of a vector and supremum norm of a bounded function. We use $\lfloor a \rfloor$ to denote the maximum integer that does not exceed a . The notation $a \lesssim b$ is used throughout to represent $a \leq cb$ for some constant c that is universal or unimportant for the analysis. Whenever possible, we use Π to represent the prior/posterior probability measure, \mathbb{P}_0 and \mathbb{E}_0 to denote the probability and expectation with respect to the distribution f_0 , and p to denote all density functions in the model except f_0 , f , and $\{f_F : F \in \mathcal{M}(\mathbb{R}^p \times \mathcal{S})\}$. For random variables, we slightly abuse the notation and do not distinguish between the random variables themselves and their realizations. We shall also use $p(x)$ or $p_x(x)$ to denote the density of the random variable x .

A weak neighborhood of f_0 is a set of densities containing a set of the form

$$V = \left\{ f \in \mathcal{M}(\mathbb{R}^p) : \left| \int \varphi_i f_0 - \int \varphi_i f \right| < \epsilon, i = 1, \dots, I \right\},$$

where φ_i 's are bounded continuous functions on \mathbb{R}^p ([Ghosal et al., 1999](#)). The posterior distribution is said to be *weakly consistent* at f_0 , if $\Pi(f \in U \mid \mathbf{y}_1, \dots, \mathbf{y}_n) \rightarrow 1$ a.s. with respect to \mathbb{P}_0 for all weak neighborhoods U of f_0 . Given a prior Π on $\mathcal{M}(\mathbb{R}^p)$, a density function $f_0 \in \mathcal{M}(\mathbb{R}^p)$ is said to be *in the KL-support* of Π , or has the *KL-property* (with

respect to Π), if $\Pi(f \in \mathcal{M}(\mathbb{R}^p) : D(f_0 \| f) < \epsilon) > 0$ for all $\epsilon > 0$. The posterior distribution is said to be L_1 (strongly) consistent at f_0 , if for all $\epsilon > 0$, $\Pi(f \in \mathcal{M}(\mathbb{R}^p) : \|f - f_0\|_1 > \epsilon \mid \mathbf{y}_1, \dots, \mathbf{y}_n) \rightarrow 0$ as $n \rightarrow \infty$ or in \mathbb{P}_0 -probability. The *posterior contraction rate* is any sequence $(\epsilon_n)_{n=1}^\infty$ such that $\Pi(f \in \mathcal{M}(\mathbb{R}^p) : \|f - f_0\|_1 > \epsilon_n \mid \mathbf{y}_1, \dots, \mathbf{y}_n) \rightarrow 0$ as $n \rightarrow \infty$ in \mathbb{P}_0 -probability. Given a family of densities \mathcal{F} on \mathbb{R}^p with a metric d on \mathcal{F} , the ϵ -covering number of \mathcal{F} with respect to d , denoted by $\mathcal{N}(\epsilon, \mathcal{F}, d)$, is defined to be the minimum number of ϵ balls of the form $\{g \in \mathcal{F} : d(f, g) < \epsilon\}$ that are needed to cover \mathcal{F} . The d -metric entropy is the logarithm of the covering number under the d -metric.

Above all, we assume that $f \sim \text{RGM}_r(\beta; g, p_\mu, p_\Sigma, p_K)$, $r = 1$ or 2 . In order to develop the posterior convergence theory, we need some regularity conditions, most of which are typically satisfied in practice. We group these conditions into two categories. The first set of conditions are the requirements for the model.

A0 The data generating density f_0 is of the form $f_0 = \phi_\Sigma * F_0$ for some $F_0 \in \mathcal{M}(\mathbb{R}^p \times \mathcal{S})$

that has a sub-Gaussian tail: $F_0(\|\mu\| \geq t) \leq B_1 \exp(-b_1 t^2)$ for some $B_1, b_1 > 0$.

A1 For some $\delta > 0, c_2 > 0$, we have $g(x) \geq c_2 \epsilon$ whenever $x \geq \epsilon$ and $\epsilon \in (0, \delta)$.

A2 g satisfies $\iint_{\mathbb{R}^p \times \mathbb{R}^p} [\log g(\|\mu_1 - \mu_2\|)]^2 p(\mu_1) p(\mu_2) d\mu_1 d\mu_2 < \infty$.

A3 For some $\underline{\sigma}^2, \bar{\sigma}^2 \in (0, +\infty)$, we have $\underline{\sigma}^2 \leq \inf_{\mathcal{S}} \lambda(\Sigma) \leq \sup_{\mathcal{S}} \lambda(\Sigma) \leq \bar{\sigma}^2$.

A4 For some (non-random) unitary $U \in \mathbb{R}^{p \times p}$, $U^T \Sigma U$ is diagonal for all $\Sigma \in \mathcal{S}$.

Condition A2 guarantees that $1/Z_K$ does not grow super-exponentially in K by **Theorem 1**. Conditions A0 and A3 guarantee that f_0 and f are not too “spiky” such that a faster rate of convergence is obtainable. Condition A4, the simultaneous diagonalizability of all $\Sigma \in \mathcal{S}$, appears to be of less importance, but it turns out that a structured space \mathcal{S} of covariance matrices decreases the $\|\cdot\|_1$ -metric entropy of the proposed sieves in Section 3.2, and hence affects the posterior contraction rate. We assume that $U^T \Sigma U = \text{diag}(\lambda_1, \dots, \lambda_p)$ for all $\Sigma \in \mathcal{S}$, *i.e.* the eigenvalues of $\Sigma \in \mathcal{S}$ are ordered according to the orthonormal eigenvectors in U .

We also need some requirements for the prior distributions.

B1 $(w_1, \dots, w_K \mid K) \sim \mathcal{D}_K(\beta)$ is weakly informative: $\beta \in (0, 1]$.

B2 $p_{\boldsymbol{\mu}}$ has a sub-Gaussian tail: $\int_{\{\|\boldsymbol{\mu}\| \geq t\}} p(\boldsymbol{\mu}) d\boldsymbol{\mu} \leq B_2 \exp(-b_2 t^2)$ for some $B_2, b_2 > 0$.

B3 For all $\boldsymbol{\mu} \in \mathbb{R}^p$, $p(\boldsymbol{\mu}) \geq B_3 \exp(-b_3 \|\boldsymbol{\mu}\|^\alpha)$ for some $\alpha \geq 2, B_3, b_3 > 0$.

B4 $p(\boldsymbol{\Sigma})$ is induced by $\prod_{j=1}^p p_\lambda(\lambda_j(\boldsymbol{\Sigma}))$ with $\text{supp}(p_\lambda) = [\underline{\sigma}^2, \bar{\sigma}^2]$.

B5 There exists some $B_4, b_4 > 0$ such that for sufficiently large K , we have

$$p_K(K) \geq \exp(-b_4 K \log K), \quad \sum_{N=K}^{\infty} p_K(N) \leq \exp(-B_4 K \log K).$$

Condition B1 assumes a vague prior on (w_1, \dots, w_K) . Conditions B2 and B3 are requirements for the tail behavior of the function $p_{\boldsymbol{\mu}}$ in the sense that they are neither heavier than Gaussian nor thinner than an exponential power density (Scricciolo et al., 2011). Condition B4 is adopted in Ghosal and Van Der Vaart (2001) to obtain an “almost” parametric convergence rate. We will also discuss possible extensions to the case where p_λ has full support on $(0, +\infty)$ later in this section. Condition B5 is the requirement for the tail behavior of the prior on K . Similar assumption on the tail behavior of the prior on K is adopted in Kruijer et al. (2010) and Shen et al. (2013) for finite mixture models. As a useful example, we show that the commonly used zero-truncated Poisson prior on K satisfies condition B5.

Example. The zero-truncated Poisson prior has a density function $p_K(K) = \frac{\mathbb{I}(K \geq 1)}{(e^\lambda - 1)K!}$ with respect to the counting measure on \mathbb{N}_+ for some intensity parameter $\lambda > 0$. Directly compute

$$\sum_{N=K+1}^{\infty} p_K(N) = \frac{1}{e^\lambda - 1} \left(e^\lambda - \sum_{N=0}^K \frac{\lambda^N}{N!} \right) = \frac{1}{e^\lambda - 1} \int_0^\lambda \frac{(\lambda - t)^K e^t dt}{K!} \lesssim \frac{\lambda^{K+1}}{(K+1)!},$$

where the second equality is due to Taylor’s expansion. By Stirling’s formula, this is further upper bounded by $(\frac{\lambda e}{K+1})^{K+1}$. Therefore, substituting $K+1$ with K , we obtain

$$\sum_{N=K}^{\infty} p_K(N) \lesssim \exp(K \log(\lambda e) - K \log K) \leq \exp\left(-\frac{1}{2} K \log K\right)$$

for sufficiently large K . The constant for \lesssim can be absorbed into the exponent, and hence

we conclude $\sum_{N=K}^{\infty} p_K(N) \leq \exp(-B_4 K \log K)$ for some $B_4 > 0$.

For the lower bound on $p(K)$, for sufficiently large K we again use Stirling's formula,

$$p(K) = \frac{1}{e^\lambda - 1} \frac{\lambda^K}{K!} \geq \exp(K \log(\lambda e) - \log K - K \log K) \geq \exp(-2K \log K).$$

Hence the zero-truncated Poisson prior on K satisfies condition B5.

3.2 Posterior Consistency

Weak consistency. Using the result from [Schwartz \(1965\)](#), a sufficient condition for Π to be weakly consistent at f_0 is that f_0 is in the KL-support of Π . The following lemma is useful in that it provides a compactly supported F_m such that f_{F_m} can approximate f_0 arbitrarily well in the KL divergence sense.

Lemma 1. *Assume conditions A0-A4 and B1-B5 hold. For all $m \in \mathbb{N}_+$, define a sequence of distributions $(F_m)_{m=1}^{\infty}$ by $F_m(A) = c_m F_0(A \cap \mathcal{T}_m)$ for any measurable $A \subset \mathbb{R}^p \times \mathcal{S}$, where*

$$\mathcal{T}_m = \left\{ (\boldsymbol{\mu} : \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu}\| \leq m, \sigma^2 + \frac{1}{m} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq \bar{\sigma}^2 - \frac{1}{m} \right\}$$

and c_m is the normalizing constant for F_m with $c_m^{-1} = F_0(\mathcal{T}_m)$. Then $\int f_0 \log \frac{f_0}{f_{F_m}} \rightarrow 0$ as $m \rightarrow \infty$.

Based on **Lemma 1**, we are able to establish the weak consistency via the KL-property.

Theorem 2. *Assume conditions A0-A4 and B1-B5 hold. Then f_0 is in the KL-support of Π , and hence $\Pi(\cdot \mid \mathbf{y}_1, \dots, \mathbf{y}_n)$ is weakly consistent at f_0 .*

Strong consistency. To establish the posterior strong consistency, we consider the following submodels of $\mathcal{M}(\mathbb{R}^p)$:

$$\mathcal{F}_{K_n} = \left\{ f_F : F = \sum_{k=1}^K w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}, K \leq K_n, \boldsymbol{\mu}_k \in \mathbb{R}^p, \boldsymbol{\Sigma}_k \in \mathcal{S} \right\}$$

and the following partition of the submodel \mathcal{F}_{K_n}

$$\mathcal{G}_K(\mathbf{a}_K) = \mathcal{F}_K \left(\prod_{k=1}^K (a_k, a_k + 1] \right), \quad \mathbf{a}_K = (a_1, \dots, a_K) \in \mathbb{N}^K, \quad K = 1, \dots, K_n,$$

where

$$\mathcal{F}_K \left(\prod_{k=1}^K (a_k, b_k] \right) = \left\{ f_F : F = \sum_{k=1}^K w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}, \|\boldsymbol{\mu}_k\|_\infty \in (a_k, b_k] \right\}.$$

The goal is to verify the following sufficient conditions for the strong consistency (Canale et al., 2017): f_0 is in the KL-support of Π , and there exists some $b, \tilde{b} > 0$, some sequence $(K_n)_{n=1}^\infty$, such that $\Pi(\mathcal{F}_{K_n}^c) \lesssim e^{-bn}$ for sufficiently large n , and for all $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} e^{-(4-\tilde{b})n\epsilon^2} \sum_{K=1}^{K_n} \sum_{a_1=0}^{\infty} \dots \sum_{a_K=0}^{\infty} \sqrt{\mathcal{N}(\epsilon, \mathcal{G}_K(\mathbf{a}_K), \|\cdot\|_1)} \sqrt{\Pi(\mathcal{G}_K(\mathbf{a}_K))} = 0. \quad (3.1)$$

Lemma 2. *Let $a_k < b_k$ be non-negative integers, $k = 1, \dots, K$. Then for sufficiently small $\delta > 0$, there exists constant $c_3 > 0$ such that*

$$\mathcal{N} \left(\delta, \mathcal{F}_K \left(\prod_{k=1}^K (a_k, b_k] \right), \|\cdot\|_1 \right) \leq \left(\frac{c_3}{\delta^{2p+1}} \right)^K \left(\prod_{k=1}^K b_k \right)^p.$$

Lemma 3. *Assume conditions A0-A4 and B1-B5 hold. Then we have*

$$\sum_{K=1}^{K_n} \sum_{a_1=0}^{\infty} \dots \sum_{a_K=0}^{\infty} \sqrt{\mathcal{N}(\delta, \mathcal{G}_K(\mathbf{a}_K), \|\cdot\|_1)} \sqrt{\Pi(\mathcal{G}_K(\mathbf{a}_K))} \leq K_n \left(\frac{M}{\delta^{p+\frac{1}{2}}} \right)^{K_n}.$$

for sufficiently small δ for some constant $M > 0$.

Based on **Lemma 2** and **Lemma 3**, we are able to verify (3.1) and hence establish the strong consistency.

Theorem 3. *Assume conditions A0-A4 and B1-B5 hold. Then $\Pi(\cdot \mid \mathbf{y}_1, \dots, \mathbf{y}_n)$ is strongly consistent at f_0 .*

3.3 Posterior Contraction Rate

To compute the posterior contraction rate, it is sufficient to find two sequences $(\underline{\epsilon}_n)_{n=1}^\infty, (\bar{\epsilon}_n)_{n=1}^\infty$ such that

$$\Pi(\mathcal{F}_n^c) \lesssim \exp(-4n\underline{\epsilon}_n^2), \quad (3.2)$$

$$\exp(-n\bar{\epsilon}_n^2) \sum_{K=1}^{K_n} \sum_{a_1=0}^{\infty} \cdots \sum_{a_K=0}^{\infty} \sqrt{\mathcal{N}(\bar{\epsilon}_n, \mathcal{G}_K(\mathbf{a}_K), \|\cdot\|_1)} \sqrt{\Pi(\mathcal{G}_K(\mathbf{a}_K))} \rightarrow 0, \quad (3.3)$$

$$\Pi\left(f : \int f_0 \log \frac{f_0}{f} \leq \underline{\epsilon}_n^2, \int f_0 \left(\log \frac{f_0}{f}\right)^2 \leq \underline{\epsilon}_n^2\right) \geq \exp(-n\underline{\epsilon}_n^2). \quad (3.4)$$

(See Theorem 3 in [Kruijer et al., 2010](#), which is also provided in Section A in the Supplementary Material). For notation convenience we refer to the set of densities $\left(f : \int f_0 \log \frac{f_0}{f} \leq \epsilon_n^2, \int f_0 \left(\log \frac{f_0}{f}\right)^2 \leq \epsilon^2\right)$ as the KL-type ball, and denote it as $B(f_0, \epsilon)$. Equation (3.4) is also known as the prior concentration condition.

Lemma 3 not only plays a fundamental role in establishing the posterior strong consistency, but also provides an upper bound for the sum in terms of δ , which is again used to verify equation (3.3). **Proposition 1** finds the rates $(\underline{\epsilon}_n)_{n=1}^\infty, (\bar{\epsilon}_n)_{n=1}^\infty$ that satisfy (3.2) and (3.3).

Proposition 1. *Assume conditions A0-A4 and B1-B5 hold. Let $\underline{\epsilon}_n = (\log n)^{t_0}/\sqrt{n}$, $\bar{\epsilon}_n = (\log n)^t/\sqrt{n}$ where t and t_0 satisfy $t > t_0 + \frac{1}{2} > \frac{1}{2}$, and $K_n = \lfloor (p+1)^{-1}(\log n)^{2t-1} \rfloor$. Then (3.2) and (3.3) hold.*

We are now left with finding the prior concentration rate $(\underline{\epsilon}_n)_{n=1}^\infty$ that satisfies (3.4). In particular, we need to bound the KL-type balls $B(f_0, \epsilon)$ by the L_1 distance. The strategy is to approximate F_0 using a finitely discrete distribution with sufficiently small number of support points. **Lemma 4** allows us to formalize this idea.

Lemma 4. *Assume conditions A0-A4 and B1-B5 hold. For some constant $\eta > 0$ and for all sufficiently small $\epsilon > 0$, there exists a discrete distribution $F^* = \sum_{k=1}^N w_k^* \delta_{(\mu_k^*, \Sigma_k^*)}$ supported*

on a subset of $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu}\|_\infty \leq 2a\}$ with $a = b_1^{-\frac{1}{2}} (\log \frac{1}{\epsilon})^{\frac{1}{2}}$, $\|\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_{k'}^*\|_\infty \geq 2\epsilon$, $|\lambda_j(\boldsymbol{\Sigma}_k^*) - \lambda_j(\boldsymbol{\Sigma}_{k'}^*)| \geq 2\epsilon$ whenever $k \neq k'$, $j = 1, \dots, p$, $N \lesssim (\log \frac{1}{\epsilon})^{2p}$, such that

$$\left\{ f_F : F = \sum_{k=1}^N w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} : (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in E_k, \sum_{k=1}^N |w_k - w_k^*| < \epsilon \right\} \subset B \left(f_0, \eta \epsilon^{\frac{1}{2}} \left(\log \frac{1}{\epsilon} \right)^{\frac{p+4}{4}} \right),$$

where

$$E_k = \left\{ (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu} - \boldsymbol{\mu}_k^*\|_\infty < \frac{\epsilon}{2}, |\lambda_j(\boldsymbol{\Sigma}) - \lambda_j(\boldsymbol{\Sigma}_k^*)| < \frac{\epsilon}{2}, j = 1, \dots, p \right\}.$$

We are in a position to derive the posterior contraction rates for the RGM model.

Theorem 4. Assume conditions A0-A4 and B1-B5 hold. Then the posterior distribution $\Pi(\cdot \mid \mathbf{y}_1, \dots, \mathbf{y}_n) \rightarrow 0$ contracts at f_0 with rate $\epsilon_n = (\log n)^t / \sqrt{n}$, $t > p + \frac{\alpha+2}{4}$.

Remark 1. Notice that the optimal rate $(\log n)^{(p+1)+} / \sqrt{n}$ is achieved when $\alpha = 2$, where $(p+1)+$ means that any $t > p+1$ is satisfied. Namely, the posterior contraction rate is optimal when p_μ has a Gaussian tail. For comparison, recall that for general location-scale Gaussian mixture problem with bounded variance, Theorem 6.2 in Ghosal and Van Der Vaart (2001) gives a contraction rate of $(\log n)^{3.5} / \sqrt{n}$ in the univariate case ($p = 1$) using the DP mixture model, in which the distribution of the location parameters is Gaussian. Analogously, in the RGM model, we may use Gaussian p_μ to control the tail rate of the joint distribution of $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)$ as $\|\boldsymbol{\mu}_k\|$ gets large, since the repulsive function h_K is bounded. **Theorem 4** improves the contraction rate to $(\log n)^t / \sqrt{n}$ with $t > 2$ compared to that given by Ghosal and Van Der Vaart (2001).

Remark 2. The boundedness on the eigenvalues of the covariance matrices (condition A3) was originally adopted in Ghosal and Van Der Vaart (2001), which is necessary to obtain an “almost” parametric rate $(\log n)^t / \sqrt{n}$ for some $t > 0$. Walker et al. (2007) adopted the same assumption and improved the posterior contraction rate of the location mixture problem. Requiring p_λ to have full support on $(0, +\infty)$, however, is necessary in cases where the underlying true density f_0 is no longer of the form $f_0 = \phi_\Sigma * F_0$ for some $F_0 \in \mathcal{M}(\mathbb{R}^p \times \mathcal{S})$.

For general mixtures of finite location mixture models, the contraction rate is known to be $(\log n)^t n^{-\tilde{\beta}/(2\tilde{\beta}+d)}$ for some $t > 0$, where f_0 is in a locally $\tilde{\beta}$ -Hölder class (Shen et al., 2013). It will be interesting to extend **Theorem 4** to the case where $\text{supp}(p_\lambda) = (0, +\infty)$ and explore the corresponding posterior contraction rate.

3.4 Model Complexity: Posterior of K

The asymptotic behavior of the posterior of K is of great interest, since it is a measurement of the complexity of a nonparametric density estimator. If a parametric assumption on f_0 is made in the sense that $f_0 = \phi_{\Sigma} * F_0$ for some finitely discrete $F_0 \in \mathcal{M}(\mathbb{R} \times \mathcal{S})$, then under mild regularity condition, Nobile (1994) proved that the posterior distribution $p(K \mid \mathbf{y}_1, \dots, \mathbf{y}_n)$ converges weakly to the point mass at K_0 a.s. under the MFM model, where K_0 is the number of support points of F_0 . However, when F_0 is no longer assumed to be finitely discrete, and a repulsive prior is introduced among components in MFM, there is little result concerning the mixture complexity in the literature. This issue is addressed in **Theorem 5** in terms of the tail behavior of the posterior of K .

Theorem 5. *Assume conditions A0-A4 and B1-B5 hold. If $p_{\mu} = \phi(\cdot \mid \mathbf{0}, \tau^2 \mathbf{I})$, then there exists some constant $\gamma > 0$, such that the tail probability of the posterior distribution of K satisfies the following inequality:*

$$\mathbb{E}_0 [\Pi(K \geq N \mid \mathbf{y}_1, \dots, \mathbf{y}_n)] \leq \exp \left(-\frac{B_4}{2} N \log N + \gamma n \right) \quad (3.5)$$

for sufficiently large N .

Corollary 1. *Assume conditions A0-A4 and B1-B5 hold. If $p_{\mu} = \phi(\cdot \mid \mathbf{0}, \tau^2 \mathbf{I})$ and the sequence $(K_n)_{n=1}^{\infty} \subset \mathbb{N}_+$ satisfies $\liminf_{n \rightarrow \infty} K_n/n > 0$, then the tail probability of the posterior of K satisfies*

$$\Pi(K \geq K_n \mid \mathbf{y}_1, \dots, \mathbf{y}_n) \rightarrow 0 \quad (3.6)$$

in \mathbb{P}_0 -probability as $n \rightarrow \infty$.

Remark 3. In terms of K , the number of support points in the RGM model, which is a measurement of the model complexity of estimating an unknown density, **Corollary 1** says that the posterior probability that K is at least a non-negligible fraction of n (in the limit) converges to 0 in \mathbb{P}_0 -probability as $n \rightarrow \infty$. In other words, the posterior number of components grows sub-linearly with respect to the sample size.

4 Posterior Inference

For the DPP mixture model, [Xu et al. \(2016\)](#) developed a variation of the RJ-MCMC sampler that can be extended to the RGM model. However, the reversible-jump moves in multi-dimensional problems could be challenging and inefficient. In this section, we design an efficient and easy-to-implement blocked-collapsed Gibbs sampler by representing the RGM model using the random partition distribution.

Let us begin with characterizing the RGM model using the latent cluster configurations. Given a random measure $F = \sum_{k=1}^K w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$ with $(w_1, \dots, w_K) \sim \mathcal{D}_K(\beta)$, we may represent the finite mixture model as follows by integrating out (w_1, \dots, w_K) :

$$\begin{aligned} (\mathbf{y}_i \mid z_i, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K, K) &\sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}), \\ p(z_1, \dots, z_n \mid K) &= \frac{\Gamma(K\beta)}{\Gamma(n + K\beta)} \prod_{k=1}^K \frac{\Gamma(\beta + \sum_{i=1}^n \mathbb{I}(z_i = k))}{\Gamma(\beta)}. \end{aligned} \quad (4.1)$$

Let \mathcal{C}_n denote the partition of $\{1, \dots, n\}$ induced by $\mathbf{z} = (z_1, \dots, z_n)$ as $\mathcal{C}_n = \{E_k : |E_k| > 0\}$ where $E_k = \{i : z_i = k\}$ for $k = 1, \dots, K$, and $|E|$ denotes the cardinality of a finite set E . For example, if one has $\mathbf{z} = (z_1, z_2, z_3, z_4, z_5, z_6) = (1, 3, 4, 4, 3, 1)$ with $n = 6$, then the corresponding partition is $\mathcal{C}_6 = \{\{1, 6\}, \{2, 5\}, \{3, 4\}\}$. Using the exchangeable partition distribution in [Miller and Harrison \(2016\)](#), we establish the generalized urn-model induced by the RGM model in **Theorem 6** after marginalizing out the intractable random distribution

F. The proof is provided in Section G of the Supplementary Material.

Theorem 6. Suppose the prior Π on $\mathcal{M}(\mathbb{R}^p)$ is defined as in Section 2, and the latent class configuration variables $\mathbf{z} = (z_1, \dots, z_n)$ is defined as in (4.1). Let $\gamma_i = \mu_{z_i}, \Gamma_i = \Sigma_{z_i}, \theta_i = (\gamma_i, \Gamma_i), i = 1, \dots, n, \mathcal{C}_{n-1}$ be the partition on $\{1, \dots, n-1\}$ induced by $\theta_1, \dots, \theta_{n-1}, (\gamma_c^* : c \in \mathcal{C}_{n-1})$ be the unique values of $(\gamma_1, \dots, \gamma_{n-1})$, and $(\Sigma_c^* : c \in \mathcal{C}_{n-1})$ be those of $(\Gamma_1, \dots, \Gamma_{n-1})$. Let $\ell = |\mathcal{C}_{n-1}|$ be the number of clusters, and K be the number of components in F , where $K \geq \ell$. Denote $\mathcal{C}_\emptyset \subset \mathbb{N}_+$ the indexes for the components associated with no observations with $|\mathcal{C}_\emptyset| = K - \ell, ((\gamma_c^*, \Gamma_c^*) \in \mathbb{R}^p \times \mathcal{S} : c \in \mathcal{C}_\emptyset)$ the component-specific parameters of the components that are not associated with any observation, and $\underline{c} = \min(c : c \in \mathcal{C}_\emptyset)$ provided that $K \geq \ell + 1$. Denote $\Pi(\theta_n \in \cdot \mid -)$ the full conditional distribution of θ_n with F marginalized out. Then for any measurable $A \subset \mathbb{R}^p \times \mathcal{S}$,

$$\Pi(\theta_n \in A \mid -) \propto \left[\frac{V_n(\ell + 1)\beta}{V_n(\ell)} \right] \sum_{K=\ell+1}^{\infty} \alpha_K G_K(A) + \sum_{c \in \mathcal{C}_{n-1}} (|c| + \beta) \phi(\mathbf{y}_n \mid \gamma_c^*, \Gamma_c^*) \delta_{(\gamma_c^*, \Gamma_c^*)}(A),$$

where

$$\begin{aligned} V_n(\ell) &= \sum_{K=\ell}^{\infty} \frac{K(K-1)\dots(K-\ell+1)}{(\beta K)(\beta K+1)\dots(\beta K+n-1)} p_K(K), \\ \alpha_K &= m_K p(K \mid \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}), \\ m_K &= \frac{\int \dots \int \phi(\mathbf{y}_n \mid \gamma_{\underline{c}}^*, \Gamma_{\underline{c}}^*) h_K(\gamma_c^* : c \in \mathcal{C}_{n-1} \cup \mathcal{C}_\emptyset) p_{\Sigma}(\Gamma_{\underline{c}}^*) d\Gamma_{\underline{c}}^* \prod_{c \in \mathcal{C}_\emptyset} p_{\mu}(\gamma_c^*) d\gamma_c^*}{\int \dots \int h_K(\gamma_c^* : c \in \mathcal{C}_{n-1} \cup \mathcal{C}_\emptyset) \prod_{c \in \mathcal{C}_\emptyset} p_{\mu}(\gamma_c^*) d\gamma_c^*}, \\ G_K(A) &\propto \iint_A L_K(\gamma_{\underline{c}}^*) \phi(\mathbf{y}_n \mid \gamma_{\underline{c}}^*, \Gamma_{\underline{c}}^*) p_{\mu}(\gamma_{\underline{c}}^*) p_{\Sigma}(\Gamma_{\underline{c}}^*) d\gamma_{\underline{c}}^* d\Gamma_{\underline{c}}^*, \\ L_K(\gamma_{\underline{c}}^*) &= \int \dots \int h_K(\gamma_c^* : c \in \mathcal{C}_{n-1} \cup \mathcal{C}_\emptyset) \prod_{c \in \mathcal{C}_\emptyset, c \neq \underline{c}} p_{\mu}(\gamma_c^*) d\gamma_c^*, \end{aligned}$$

and $h_K(\gamma_c : c \in \mathcal{C}_{n-1} \cup \mathcal{C}_\emptyset) = h_K(\gamma_{c_1}^*, \dots, \gamma_{c_K}^*)$ if one labels $\mathcal{C}_{n-1} \cup \mathcal{C}_\emptyset$ as $\{c_1, \dots, c_K\}$.

Theorem 6 is instructive for deriving the blocked-collapsed Gibbs sampler for the posterior inference on \mathcal{C} and $(\theta_c^* : c \in \mathcal{C})$. We follow the notation in **Theorem 6**. Let \mathcal{C}_{-i} be

the partition induced by $\boldsymbol{\theta}_{-i} := (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) \setminus \{\boldsymbol{\theta}_i\}$. Notice that by exchangeability

$$\begin{aligned}\Pi(\mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\} \mid \mathbf{y}_i, \boldsymbol{\theta}_{-i}) &\propto \left(\frac{V_n(|\mathcal{C}_{-i}| + 1)}{V_n(|\mathcal{C}_{-i}|)} \beta \right) \sum_{K=|\mathcal{C}_{-i}|+1}^{\infty} \alpha_K, \\ \Pi(\mathcal{C} = (\mathcal{C}_{-i} \setminus \{c\}) \cup \{c \cup \{i\}\} \mid \mathbf{y}_i, \boldsymbol{\theta}_{-i}) &\propto \phi(\mathbf{y}_i \mid \boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*) (|c| + \beta),\end{aligned}$$

where $c \in \mathcal{C}_{-i}$. Namely, given a partition \mathcal{C}_{-i} on $\{1, \dots, n\} \setminus \{i\}$, the left-out index i forms a new singleton cluster with probability proportional to $\left[\frac{V_n(|\mathcal{C}_{-i}| + 1)}{V_n(|\mathcal{C}_{-i}|)} \beta \right] \sum_{K=|\mathcal{C}_{-i}|+1}^{\infty} \alpha_K$, and is merged into an existing cluster $c \in \mathcal{C}_{-i}$ with probability proportional to $\phi(\mathbf{y}_i \mid \boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*) (|c| + \beta)$. Instead of directly sampling from above categorical distribution, which involves computing the intractable α_K 's, we take advantage of the integral structure of α_K and design auxiliary variables.

Suppose the current state of the Markov chain consists of $(\boldsymbol{\theta}_c^*, c \in \mathcal{C}_n)$, and a partition \mathcal{C}_n on $\{1, \dots, n\}$. We instantiate $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)$ using $(\boldsymbol{\theta}_c^* : c \in \mathcal{C}_n)$ and \mathcal{C}_n by letting $\boldsymbol{\theta}_{z_i} = \boldsymbol{\theta}_c^*$ if $i \in c$. A complete iteration of the blocked-collapsed Gibbs sampler consists of the following steps:

• **Step 1: For $i = 1, \dots, n$:**

- i) **Sample $K \sim p(K \mid \mathcal{C}_n = \mathcal{C}_{-i} \cup \{\{i\}\})$** ; Set $\ell = |\mathcal{C}_{-i}|$, compute \mathcal{C}_{\emptyset} with $|\mathcal{C}_{\emptyset}| = K - \ell$, and set $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) \setminus \{\boldsymbol{\theta}_i\}$.
- ii) **Sample auxiliary variables $(\boldsymbol{\gamma}_c^* : c \in \mathcal{C}_{\emptyset})$ and $\boldsymbol{\Gamma}_{\underline{c}}^*$** . Sample $\boldsymbol{\Gamma}_{\underline{c}}^* \sim p_{\Sigma}(\boldsymbol{\Gamma}_{\underline{c}}^*)$. Sample $(\boldsymbol{\gamma}_c^* : c \in \mathcal{C}_{\emptyset})$ by accept-reject sampling: Sample $(\boldsymbol{\gamma}_c^* : c \in \mathcal{C}_{\emptyset}) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I})$ and $U \sim \text{Unif}(0, 1)$ independently; If $U < h_K(\boldsymbol{\gamma}_c^* : c \in \mathcal{C}_{-i} \cup \mathcal{C}_{\emptyset})$, then accept the new proposed samples; Otherwise resample $(\boldsymbol{\gamma}_c^* : c \in \mathcal{C}_{\emptyset}) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I})$ and U until $U < h_K(\boldsymbol{\gamma}_c^* : c \in \mathcal{C}_{-i} \cup \mathcal{C}_{\emptyset})$.
- iii) **Sample $\mathcal{C} \sim p(\mathcal{C} \mid -)$** . This is done by sampling from the categorical distribution

$$\begin{aligned}\mathbb{P}(\mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\} \mid -) &\propto \left[\frac{V_n(|\mathcal{C}_{-i}| + 1)}{V_n(|\mathcal{C}_{-i}|)} \beta \right] \phi(\mathbf{y}_i \mid \boldsymbol{\gamma}_{\underline{c}}^*, \boldsymbol{\Gamma}_{\underline{c}}^*), \\ \mathbb{P}(\mathcal{C} = (\mathcal{C}_{-i} \setminus \{c\}) \cup \{c \cup \{i\}\} \mid -) &\propto (|c| + \beta) \phi(\mathbf{y}_i \mid \boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*).\end{aligned}$$

If $\mathcal{C}_n = \mathcal{C}_{-i} \cup \{\{i\}\}$, and $\{\{i\}\}$ originally is a singleton, then leave θ_i unchanged;

If $\mathcal{C}_n = \mathcal{C}_{-i} \cup \{\{i\}\}$, and $\{\{i\}\}$ originally is not a singleton, then set $\theta_i = (\gamma_{\underline{c}}^*, \Gamma_{\underline{c}}^*)$;

If $\mathcal{C}_n = (\mathcal{C}_{-i} \setminus \{c\}) \cup \{c \cup \{i\}\}$ for some $c \in \mathcal{C}_{-i}$, then set $\theta_i = \theta_c^*$. Change the current state of the Markov chain to $(\theta_c^*, c \in \mathcal{C}_n)$ and \mathcal{C}_n using $(\theta_1, \dots, \theta_n)$.

- **Step 2: Sample** $K \sim p(K \mid \mathcal{C}_n)$; Set $\ell = |\mathcal{C}_n|$, and compute \mathcal{C}_\emptyset such that $|\mathcal{C}| = K - \ell$.
- **Step 3: Sample** $(\Gamma_c^* : c \in \mathcal{C}_n)$: For all $c \in \mathcal{C}$, sample Γ_c^* from

$$p(\Gamma_c^* \mid -) \propto p_\Sigma(\Sigma_c^*) \prod_{i \in c} \phi(\mathbf{y}_i \mid \gamma_c^*, \Sigma_c^*).$$

- **Step 4(Blocking): Jointly sample** $(\gamma_c^* : c \in \mathcal{C}_n)$. This can be done by accept-reject sampling: For each $c \in \mathcal{C}_n$, sample

$$p(\gamma_c^* \mid -) \propto p_\mu(\gamma_c^*) \prod_{i \in c} \phi(\mathbf{y}_i \mid \gamma_c^*, \Gamma_c^*),$$

and for each $c \in \mathcal{C}_\emptyset$, sample $\gamma_c^* \sim p_\mu(\gamma_c^*)$. Next independently sample $U \sim \text{Unif}(0, 1)$;

If $U < h_K(\gamma_c^* : c \in \mathcal{C}_n \cup \mathcal{C}_\emptyset)$, then accept the new proposed samples; Otherwise resample $(\gamma_c^* : c \in \mathcal{C}_n \cup \mathcal{C}_\emptyset)$ and U until $U < h_K(\gamma_c^* : c \in \mathcal{C}_n \cup \mathcal{C}_\emptyset)$.

- **Step 5:** Change the current state to $(\theta_c^*, c \in \mathcal{C}_n)$ and \mathcal{C}_n .

Remark 4. The proposed sampler can be easily extended to the case where a non-Gaussian mixture model is used, provided we use priors p_μ, p_Σ in (2.3) that are conjugate to the non-Gaussian kernel density. In cases where non-conjugate priors p_μ, p_Σ are used, it is also possible to extend the blocked-collapsed Gibbs sampler either by a method of “no-gaps” proposed by [MacEachern and Mueller \(1998\)](#) or a Metropolis-within-Gibbs sampler ([Neal, 2000](#)).

5 Numerical Examples

We evaluate the performance of the RGM model and the blocked-collapsed Gibbs sampler proposed in Section 4 through extensive simulation studies and real data analysis. Subsections 5.1 and 5.2 aim to illustrate the advantages of the RGM concerning accurate density estimation, identification of correct number of components, and the model complexity. Subsection 5.3 demonstrates the efficiency of the proposed blocked-collapsed Gibbs sampler compared to the DP mixture model and the DPP mixture model (Xu et al., 2016). In Subsection 5.4 we apply the RGM model to analyze the Old Faithful geyser eruption data (Silverman, 1986). We assume $\beta = 1$, indicating a uniform prior on $(w_1, \dots, w_K \mid K)$. We assign a zero-truncated Poisson prior on K with intensity $\lambda = 1$, that is, $p(K) = \frac{\mathbb{I}(K \geq 1)}{(e-1)K!}$. Without loss of generality, the repulsive function is defined as $g(x) = \frac{x}{g_0 + x}$ for some $g_0 > 0$ and h_K is of the form (2.4). Lastly, we assume $p(\boldsymbol{\mu}) = \phi(\boldsymbol{\mu} \mid 0, \tau^2 \mathbf{I}_p)$ and a truncated inverse Gamma prior on $\lambda(\boldsymbol{\Sigma})$, $p(\lambda) \propto \mathbb{I}(\underline{\sigma}^2 \leq \lambda \leq \bar{\sigma}^2) \lambda^{-a_0-1} \exp(-b_0/\lambda)$ for some $a_0, b_0 > 0$.

We give the convergence diagnostics via trace plots and autocorrelation plots in the Section I of the Supplementary Material. To compare the performance of the proposed models with the competitors (*e.g.* the DP mixture (DPM) model and the DPP mixture model), we follow the ideas in Pettit (1990) and compute the *logarithm of the conditional predictive ordinate* (log-CPO) of different models using the post-burn-in samples as follows:

$$\text{log-CPO} = - \sum_{i=1}^n \log \left[\frac{1}{n_{\text{mc}}} \sum_{i_{\text{it}}=1}^{n_{\text{mc}}} p(\mathbf{y}_i \mid \boldsymbol{\Theta}_{\text{mc}}^{i_{\text{it}}}) \right],$$

where n_{mc} is the number of the post-burn-in MCMC samples, i_{it} indexes the post-burn-in iterations, and $\boldsymbol{\Theta}_{\text{mc}}^{i_{\text{it}}}$ represent the post-burn-in samples of all parameters generated by the MCMC at the i_{it} th iteration.

5.1 Fitting Multi-modal Density: Finite Gaussian Mixtures

In this subsection, to demonstrate multi-modal density fitting, we fit a finite mixture of Gaussians using the RGM model, and evaluate its performance regarding the density estimation and the identification of the number of components. In particular, suppose the simulated data $\mathbf{y}_1, \dots, \mathbf{y}_n$, $n = 1000$, are i.i.d. generated from the bivariate density:

$$f_0(\mathbf{y}) = 0.4\phi(\mathbf{y} \mid \mathbf{0}, \text{diag}(2, 1)) + 0.3\phi(\mathbf{y} \mid (-6, -6)^T, 3\mathbf{I}_2) + 0.3\phi(\mathbf{y} \mid (6, 6)^T, 2\mathbf{I}_2).$$

We implement the proposed blocked-collapsed Gibbs sampler with $g_0 = 10$, $\tau = 1$, $m = 2$, $\underline{\sigma} = 0.1$, $\bar{\sigma} = 10$, and a total number of 2000 iterations with the first 1000 iterations discarded as burn-in. For comparison, we consider the following DPM model,

$$(\mathbf{y}_i \mid \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \sim N(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}), \quad (\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i} \mid G) \stackrel{\text{i.i.d.}}{\sim} G, \quad \text{and } (G \mid \alpha, G_0) \sim \text{DP}(\alpha, G_0),$$

where $G_0 = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \sim N(\mathbf{m}_1, \boldsymbol{\Sigma}/k_0)$ and $\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(4, \boldsymbol{\Psi}_1)$, $\alpha \sim \text{Gamma}(1, 1)$, $\mathbf{m}_1 \sim N(\mathbf{0}, 2\mathbf{I}_2)$, $k_0 \sim \text{Gamma}(0.5, 0.5)$, and $\boldsymbol{\Psi}_1 \sim \text{Inv-Wishart}(4, 0.5\mathbf{I}_2)$. For the DP mixture model, we use K to represent the number of clusters throughout this section, since the number of components is always infinity.

Table 1 shows that the log-CPO of the RGM model is higher than that of the DPM model, indicating that RGM is preferred according to the data. Figures 1a and 1c show the posterior density estimation under the RGM model and the DP mixture model, respectively, indicating that both methods perform well in terms of density estimation.

Table 1: Log-Conditional Predictive Ordinate (log-CPO) for Numerical Results

Model	Subsection 5.1	Subsection 5.2	Subsection 5.4
RGM model	-3584.567	-3462.07	-242.7644
DPM model	-4599.204	-3483.667	-315.1032
DPP mixture model			-512.6564

However, as shown in the histograms of posterior number of components/clusters in Fig-

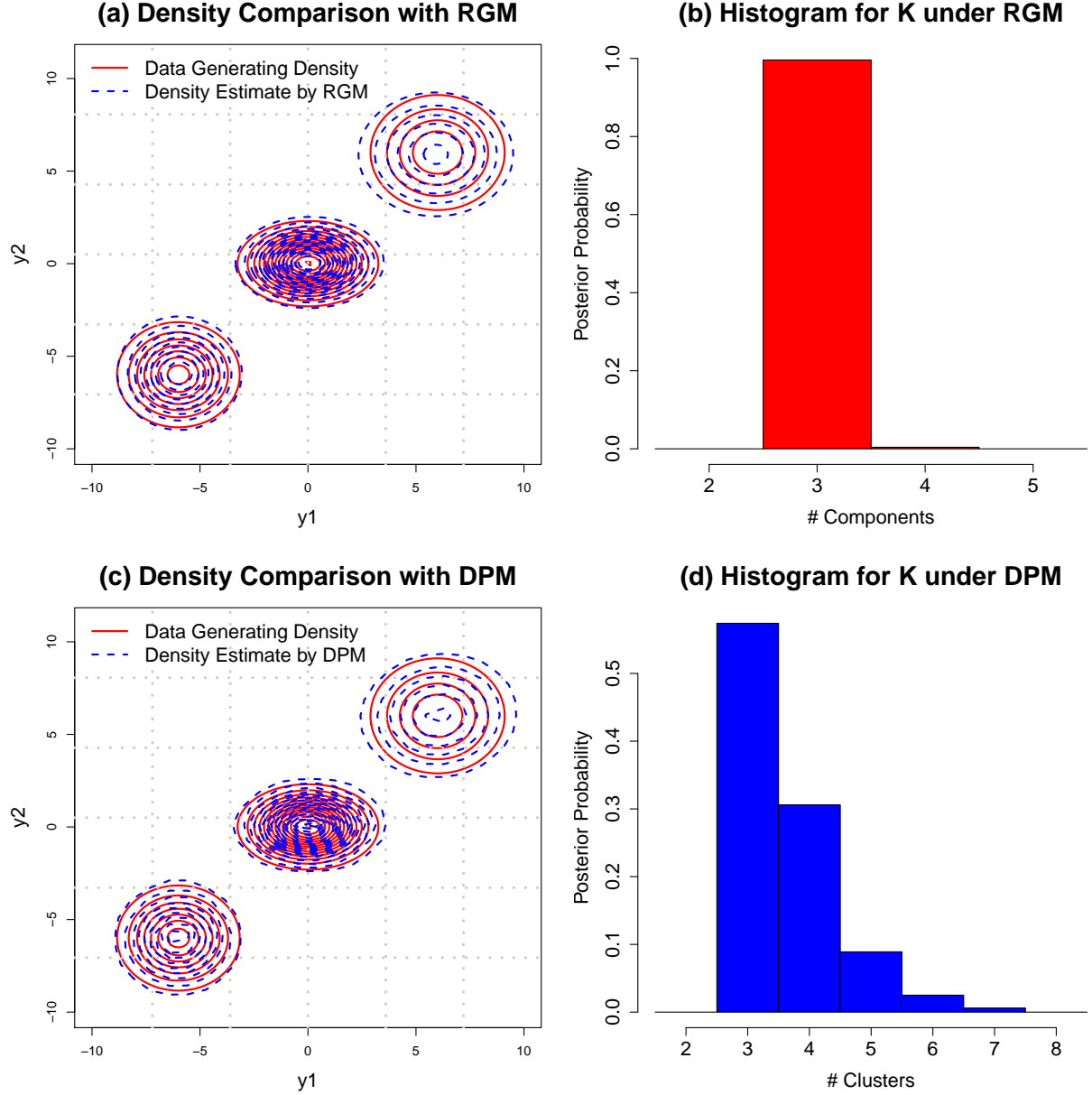


Figure 1: Fitting Multi-modal Density: Panels (a) and (c) are the contour plots for the posterior density estimation of the RGM model and the DPM model, respectively. Panels (b) and (d) are the histograms of the posterior number of components under the RGM model and the posterior number of clusters under the DPM model, respectively, where the underlying true number of components is $K = 3$.

ures 1b and 1d, the posterior distribution of the number of components is highly concentrated around the underlying true K under the RGM model, whereas the DPM model assigns relatively higher posterior probability to redundant clusters. This agrees with the inconsistency phenomenon of the DPM model for the identification of number of components, which was previously reported in Miller and Harrison (2013).

5.2 Fitting Uni-modal Density: Continuous Gaussian Mixtures

Rather than generating the simulated data from a finite discrete Gaussian mixture model, in this subsection we consider a continuous mixture of Gaussians,

$$f_0(y_1, y_2) = \prod_{j=1}^2 \int_0^\infty \phi(y_j - \mu_j - \mu_0 \mid 0, 1) \exp(-\mu_j) d\mu_j. \quad (5.1)$$

Notice that f_0 is uni-modal. The random variables y_i , $i = 1, 2$ can be i.i.d. generated as the sum of a normal random variable and an exponential random variable with intensity parameter 1, *i.e.* $y_i = z_i + \mu_i$ where $z_i \sim N(\mu_0, 1)$ and $\mu_i \sim \text{Exp}(1)$, $i = 1, 2$. Then $\mathbf{y} = (y_1, y_2)$ is the random vector following the distribution in (5.1). The marginal distribution of y_i is referred to as the *exponentially modified Gaussian* (EMG) distribution, the density of which can be alternatively represented as $f(y) = \frac{1}{2} \exp(\mu_0 - y + \frac{1}{2}) \text{erfc}\left(\frac{\mu_0 + 1 - y}{\sqrt{2}}\right)$, where erfc is the well-known complementary error function $\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \exp(-t^2) dt$. We generate $n = 1000$ i.i.d. samples from f_0 with $\mu_0 = -4$, and implement the proposed blocked-collapsed Gibbs sampler with $g_0 = 7$, $\tau = 1$, $m = 2$, $\underline{\sigma} = 0.1$, $\bar{\sigma} = 10$, and a total number of 2000 iterations with the first 1000 iterations discarded as burn-in phase. For comparison, we consider the similar DPM model with the same setting as in Section 5.1.

Figures 2a and 2c show that the RGM model and the DP mixture model provide similar accurate density estimation to the underlying true density f_0 . However, Figures 2b and 2d indicate that under the DPM model, the number of active components tends be larger than that under the RGM model in order to fit the data well. In other words, the posterior of the

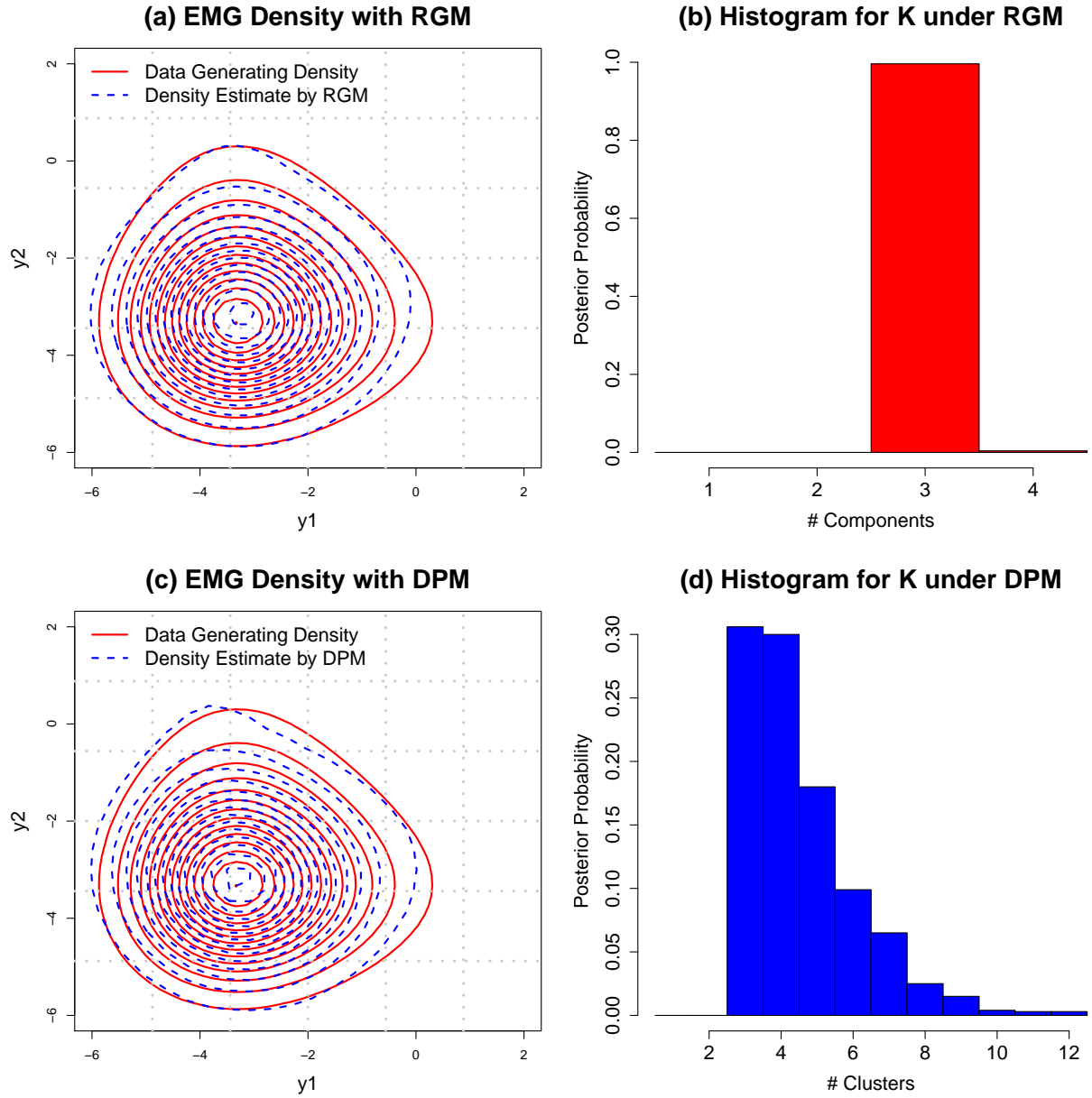


Figure 2: Fitting Uni-modal Density: Panels (a) and (c) are the contour plots for the posterior density estimation under the RGM model and the DPM model, respectively. Panels (b) and (d) are the histograms of the posterior number of components under the RGM model and the posterior number of clusters under the DPM model, respectively.

RGM provides the same level of accuracy in density estimation as the DPM model does, but with less number of components. In this simulation study, with high posterior probability, the RGM model only utilizes 3 components to fit the density, whereas DPM model assigns large posterior probability to utilizing 4 or more components. The log-CPO comparison in Table 1, clearly show that the RGM model outperforms the DPM model.

5.3 Multivariate Model-Based Clustering

Now we focus on a higher dimensional model-based clustering problem. Suppose that we generate $n = 500$ i.i.d. samples from a mixture of 3 10-dimensional Gaussians:

$$f_0(\mathbf{y}) = 0.4\phi(\mathbf{y} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.3\phi(\mathbf{y} \mid \boldsymbol{\mu}_2, 3\mathbf{I}_{10}) + 0.3\phi(\mathbf{y} \mid \boldsymbol{\mu}_3, 2\mathbf{I}_{10}),$$

where the covariance matrix for the first component is a randomly generated diagonal matrix:

$$\boldsymbol{\Sigma}_1 = \text{diag}(5.5729, 5.0110, 3.6832, 8.1931, 5.7717, 3.0267, 3.5011, 7.8291, 4.2233, 4.3885),$$

and $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = (-6, \dots, -6)^T \in \mathbb{R}^{10}$, $\boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2$. In this simulation study, we focus on the model-based clustering without fixing the number K of components *a priori*. Due to the challenge of visualizing high-dimensional clustering, we only show the scatter plot of the 4th versus 8th coordinate of the simulated data in Figure 3a. These two dimensions correspond to the first two largest eigenvalues in the covariance matrix. The projection of the data onto this 2-dimensional subspace shows that the three clusters are not well-separated. We implement the proposed blocked-collapsed Gibbs sampler with $g_0 = 70$, $\tau = 1$, $m = 2$, $\underline{\sigma} = 0.1$, $\bar{\sigma} = 10$. To demonstrate the efficiency of the proposed sampler, we keep all MCMC samples and compare the efficiency of the algorithms in terms of their numbers of burn-in iterations.

For comparison, we consider the two alternative clustering models and evaluate their performance in terms of efficiency in estimating posterior number of components. The first one is the DPM model: $(\mathbf{y}_i \mid \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \sim N(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$, $(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i} \mid G) \stackrel{\text{i.i.d.}}{\sim} G$, and $(G \mid \alpha, G_0) \sim$

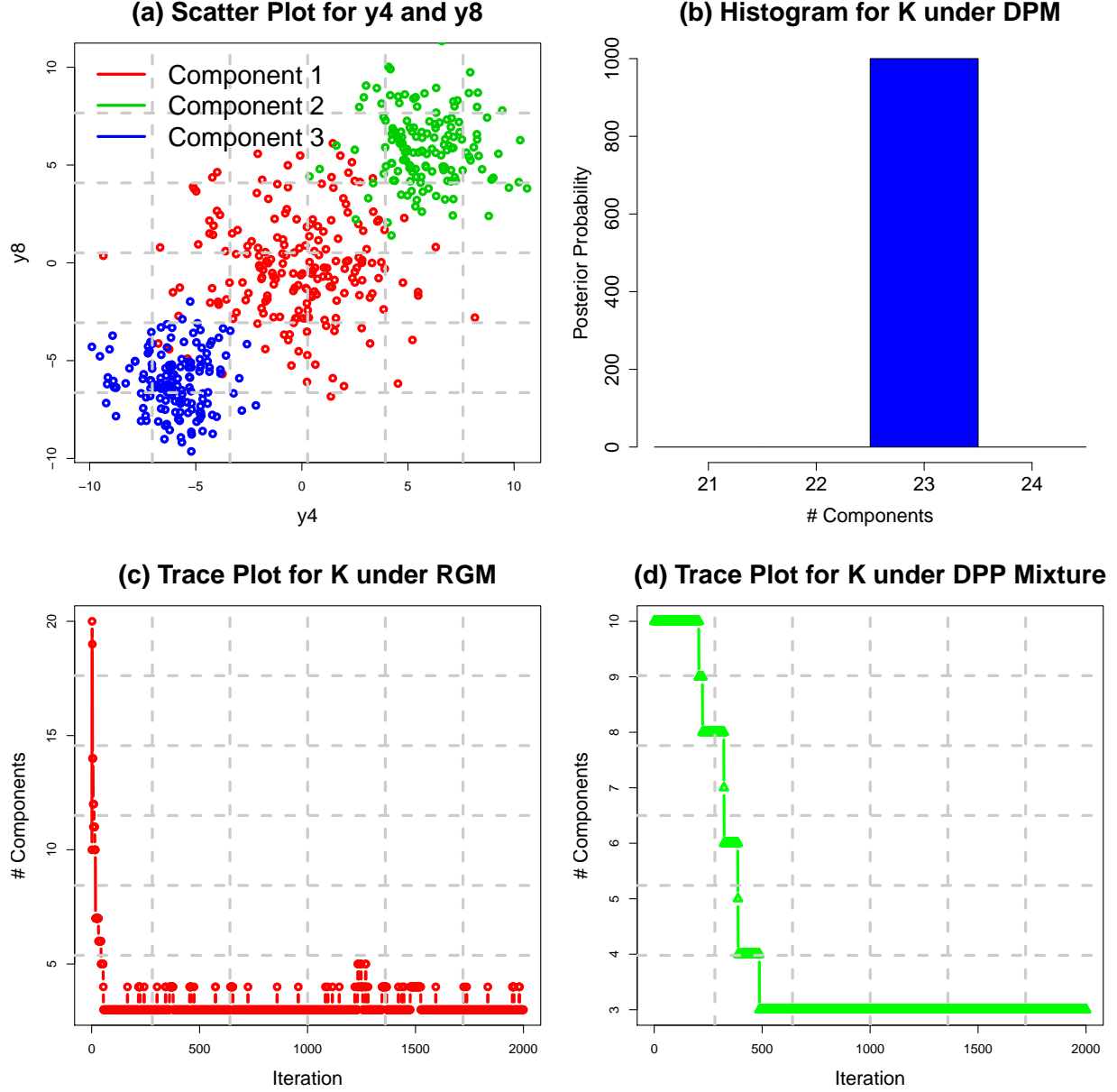


Figure 3: Multivariate Model-Based Clustering: Panel (a) is the scatter plot of the 4th-versus-8th coordinate of the simulated data; Panel (b) is the histogram of the posterior number of clusters under the DPM model; Panels (c) and (d) are the trace plots for the posterior samples of K under the RGM model, and that of the number of clusters under the DPP mixture model, respectively.

$\text{DP}(\alpha, G_0)$, where $G_0 = \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}/k_0)$ and $\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(12, \boldsymbol{\Psi}_1)$, $\alpha = 1$, $k_0 \sim \text{Gamma}(0.005, 0.005)$, and $\boldsymbol{\Psi}_1 = 0.1\mathbf{I}_{10}$. The Second alternative model is the *DPP*

mixture model proposed in [Xu et al. \(2016\)](#), who used the determinantal point process as a repulsive function: $h_K(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K) = \det \left\{ \left[\exp \left(-\frac{1}{2\theta^2} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|^2 \right) \right]_{K \times K} \right\}$ for $K \geq 2$, $h_K \equiv 1$ otherwise. The posterior inference of the DPP mixture model was performed using a potentially inefficient RJ-MCMC sampler. We initialize the Markov chains with $K = 10$ for all three models. By comparing the histogram and trace plots of the posterior number of components/clusters in Figures 3b, 3c, and 3d, we find the DPM model significantly overestimates the number of components at 23 in order to fit the 10-dimensional data well; The DPP mixture inferred with RJ-MCMC, though eventually stabilizes at the correct $K = 3$, requires relatively large number of iterations to find the underlying truth (approximately 500 iterations). In contrast, the posterior number of components under the RGM model highly concentrates around the underlying true $K = 3$, and stabilizes within only 100 iterations. In terms of efficiency of the Markov chain, the blocked-collapsed Gibbs sampler of the RGM model outperforms the other two alternatives.

We further report the performance of the model-based clustering procedure under the RGM model. Adopting the ideas in [Xu et al. \(2016\)](#) and [Dahl \(2006\)](#), we define the association matrix $S \in \{0, 1\}^{n \times n}$ with (i, j) th entries being $\mathbb{I}(\gamma_i = \gamma_j)$, and $H \in \{0, 1\}^{n \times n}$ with (i, j) th entries being $\mathbb{I}(\gamma_i = \gamma_j \mid \mathbf{y}_1, \dots, \mathbf{y}_n)$. Using the posterior samples, H can be approximated using the posterior mean of $\mathbb{I}(\gamma_i = \gamma_j)$ for all (i, j) pairs. We compute the mean of the absolute mis-classification matrix $(|H_{ij} - S_{ij}|)_{n \times n}$. The mis-classification error defined by $\frac{1}{n^2} \|\hat{H} - S\|_F$ is 1.0215×10^{-5} , where \hat{H} is computed using the posterior means.

5.4 Old Faithful Geyser Eruption Data

In this subsection, we consider the Old Faithful geyser eruption data that record the eruption length of the Old faithful geyser in the Yellowstone National Park with the number of observations $n = 272$ as a real world example. Following the procedure described in [Qin and Priebe \(2013\)](#); [Garcia-Escudero and Gordaliza \(1999\)](#) for each observed eruption duration

time, we pair it with the time length of the next eruption, so that we have a bivariate data of sample size 271. The points with the “short followed by short” eruption property were identified as outliers in [Garcia-Escudero and Gordaliza \(1999\)](#), in which a robust trimmed mean procedure was used to reduce the effects from these outliers. Alternatively, we apply the RGM model to analyze the bivariate dataset, and show that the outliers can actually be identified as an extra component. We also compare the proposed method with the two alternative models: the DPM model and the DPP mixture model as described in subsection [5.3](#).

Figure [4](#) shows the predictive densities and the histograms of the number of components/clusters estimated by the three models: the RGM model, the DPM model, and the DPP mixture model. The proposed RGM, not only identifies the outliers component (Figure [4a](#)), but also provides the posterior number of components that is highly concentrated at $K = 4$ (Figure [4b](#)). In contrast, Figure [4c](#) shows that DPP mixture fails to identify the outliers at the bottom-left corner of the scatter plot – instead, they are merged into the existing cluster located at the bottom-right corner. The corresponding posterior number of components K , as illustrated in Figure [4d](#), is highly concentrated at $K = 3$, failing to detect the outlier component. In addition, notice that failure in identifying the outliers significantly affects the posterior predictive density estimate, as shown from the comparison of the level curves among Figures [4a](#), [4c](#), and [4e](#). The DPM model in Figure [4e](#), although successfully detects the outliers component, still assigns relatively larger posterior probability to redundant components (Figure [4f](#)). Hence the proposed RGM model outperforms the other two alternatives in terms of the robustness or the model complexity measured by the posterior of K . This conclusion is also supported by the fact that log-CPO of the RGM model is higher than those of the DPM model and the DPP mixture model (Table [1](#)).

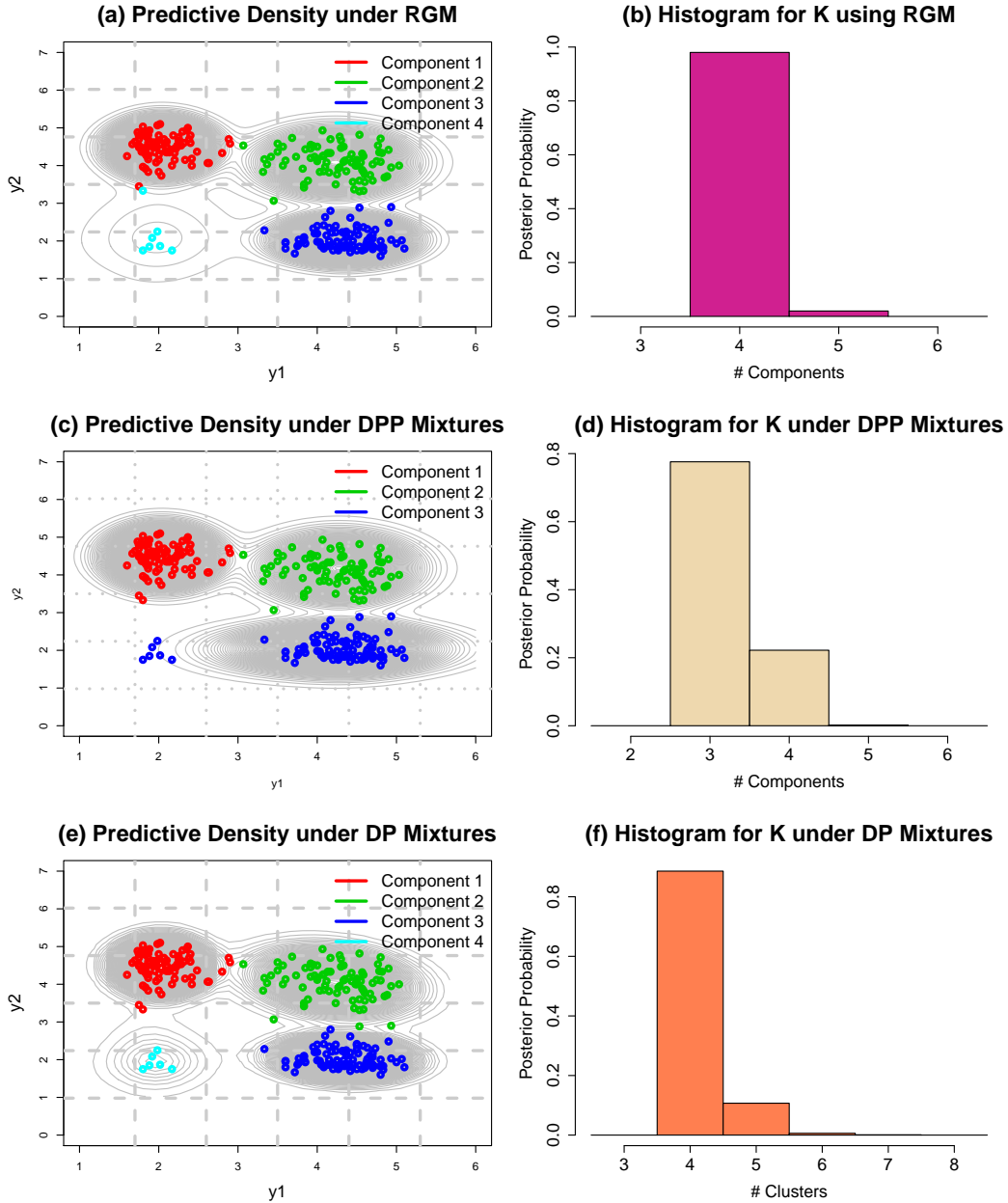


Figure 4: Old Faithful Geyser Eruption Data: Panels (a), (c), and (e) are the scatter plots of the observations with their corresponding clusters and contour plots of the posterior predictive density estimate (grey level curves) stratified by the RGM model, the DPP mixture model, and the DPM model, respectively. Panels (b), (d), and (f) are the histograms of the posterior distributions of the number of components/clusters under the RGM model, the DPP mixture model, and the DPM model, respectively.

6 Conclusion

We propose the RGM model, in which the location parameters for each component are not *a priori* independent, but jointly distributed according to some symmetric repulsive distribution that encourages the separation of the locations for different components. We establish the posterior consistency and obtain an “almost” parametric posterior contraction rate $((\log n)^t/\sqrt{n}$ with $t > p + 1$), generalizing the repulsive mixture model proposed by [Petrálie et al. \(2012\)](#); [Quinlan et al. \(2017\)](#) to the context of nonparametric density estimation. Furthermore, we study the asymptotic behavior of the model complexity of the proposed RGM model regarding the number of necessary components needed to fit the data well.

Based on the exchangeable partition distribution, we develop a blocked-collapsed Gibbs sampler for the posterior inference. Through extensive simulation studies and real data analysis, we demonstrate that the proposed RGM model is able to detect outliers and simultaneously penalize the number of components to reduce model complexity and accurately estimate the underlying true density. Moreover, the proposed sampler converges much faster than the RJ-MCMC sampler in [Xu et al. \(2016\)](#) even in slightly higher dimensional clustering problems.

There are several potential further extensions. Beyond mixture models for density estimation, it is also interesting to extend the repulsive mixture model to the nested clustering of grouped data, and perform simultaneous clustering of individuals within each group and the group level features when the inference prefers the parsimonious model and the focus is the interpretation of the clusters as meaningful subgroups. Secondly, the posterior distribution of the number of components under the RGM model is potentially sensitive to the hyperparameters in the repulsive function h_K . Performing sensitivity analysis by imposing suitable priors on these hyperparameters is possible if an efficient updating rule for them can be integrated within the blocked-collapsed Gibbs sampler. Lastly, instead of implementing

a Gibbs sampler, which is not scalable to large number of observations, one can develop an optimization-based fast inference algorithm, which would greatly improve the computational efficiency and scalability of the posterior inference.

References

- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *The Annals of Statistics*, pages 1152–1174.
- Canale, A., De Blasi, P., et al. (2017). Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, 23(1):379–404.
- Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian Inference for Gene Expression and Proteomics*, pages 201–218.
- Fuquene, J., Steel, M., and Rossell, D. (2016). On choosing mixture components via non-local priors. *arXiv preprint arXiv:1604.00314*.
- Garcia-Escudero, L. A. and Gordaliza, A. (1999). Robustness properties of k-means and trimmed k-means. *Journal of the American Statistical Association*, 94(447):956–969.
- Ghosal, S., Ghosh, J. K., Ramamoorthi, R., et al. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158.
- Ghosal, S. and Van Der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29(5):1233–1263.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

- Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.
- Ishwaran, H. and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11(3):508–532.
- Kruijer, W., Rousseau, J., Van Der Vaart, A., et al. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257.
- MacEachern, S. N. and Mueller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.
- Miller, J. W. and Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems*, pages 199–206.
- Miller, J. W. and Harrison, M. T. (2016). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, (just-accepted).
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.
- Nobile, A. (1994). *Bayesian analysis of finite mixture distributions*. PhD thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.
- Petralia, F., Rao, V., and Dunson, D. B. (2012). Repulsive mixtures. In *Advances in Neural Information Processing Systems*, pages 1889–1897.
- Pettit, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 52(1):175–184.

- Qin, Y. and Priebe, C. E. (2013). Maximum Lq-likelihood estimation via the expectation-maximization algorithm: a robust estimation of mixture models. *Journal of the American Statistical Association*, 108(503):914–928.
- Quinlan, J. J., Quintana, F. A., and Page, G. L. (2017). Parsimonious hierarchical modeling using repulsive distributions. *arXiv preprint arXiv:1701.04457*.
- Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26.
- Scricciolo, C. et al. (2011). Posterior rates of convergence for Dirichlet mixtures of exponential power densities. *Electronic Journal of Statistics*, 5:270–308.
- Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.
- Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation*®, 36(1):45–54.
- Walker, S. G., Lijoi, A., Pruenster, I., et al. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics*, 35(2):738–746.
- Wu, Y. and Ghosal, S. (2010). The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 101(10):2411–2419.
- Wu, Y., Ghosal, S., et al. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, 2:298–331.

Xu, Y., Mueller, P., and Telesca, D. (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics*, 72(3):955–964.

Bayesian Repulsive Gaussian Mixture Model

Supplementary Material

A Supporting Results

Sufficient Conditions for Posterior Weak Consistency

We use the results in [Wu et al. \(2008\)](#) to establish the weak consistency of Π . Denote Π^* the prior on $F \in \mathcal{M}(\mathbb{R}^p \times \mathcal{S})$ that induces the prior Π on f . Notice that the prior Π^* on F is supported on the class of all finitely discrete probability distributions on $\mathbb{R}^p \times \mathcal{S}$, which is dense in $\mathcal{M}(\mathbb{R}^p \times \mathcal{S})$ under the weak topology, we conclude that Π^* has the weak full support on $\mathcal{M}(\mathbb{R}^p \times \mathcal{S})$. As a consequence, we need to verify the conditions A1, A7, A8, and A9 (which we list as C1, C2, C3, and C4) there: For all $\epsilon > 0$ in [Wu et al. \(2008\)](#) exists some $F_\epsilon \in \text{supp}(\Pi^*)$, a closed set $D \supset \text{supp}(F_\epsilon)$ such that

$$\text{C1 } \int_{\mathbb{R}^p} f_0(\mathbf{y}) \log \frac{f_0(\mathbf{y})}{f_{F_\epsilon}(\mathbf{y})} d\mathbf{y} < \epsilon;$$

$$\text{C2 } \int_{\mathbb{R}^p} f_0(\mathbf{y}) \left| \log \frac{f_{F_\epsilon}(\mathbf{y})}{\inf_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in D} \phi(\mathbf{y} | \boldsymbol{\mu}, \boldsymbol{\Sigma})} \right| d\mathbf{y} < \infty;$$

$$\text{C3 } \text{For any compact } C \subset \mathbb{R}^p, c := \inf_{(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \in C \times D} \phi(\mathbf{y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) > 0;$$

$$\text{C4 } \text{For any compact } C \subset \mathbb{R}^p, \text{ there exists some } E \subset \mathbb{R}^p \times \mathcal{S} \text{ such that } D \text{ is contained in the interior of } E, \text{ the class of functions } \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mapsto \phi(\mathbf{y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) : \mathbf{y} \in C\} \text{ is uniformly equicontinuous on } E, \text{ and } \sup\{\phi(\mathbf{y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) : \mathbf{y} \in C, (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in E^c\} < c\epsilon/4.$$

Sufficient Conditions for Posterior Strong Consistency

To prove the posterior strong consistency of the RGM model we apply Theorem 1 in [Canale et al. \(2017\)](#).

Theorem A.1. *Consider a statistical \mathcal{F} with a prior Π , let $(\mathbf{y}_i)_{i=1}^n$ be an i.i.d. sequence with density $f_0 \in \mathcal{F}$. Assume that there exists a sequence of submodels $(\mathcal{F}_n)_{n=1}^\infty$ with partitions*

$\mathcal{F}_n = \bigcup_{j=1}^{\infty} \mathcal{F}_{nj}$. If f_0 is in the KL-support of Π , and there exists some $a, b > 0$ such that $\Pi(\mathcal{F}_n^c) \lesssim e^{-bn}$, and

$$\exp(-(4-a)n\epsilon^2) \sum_{j=1}^{\infty} \sqrt{\mathcal{N}(2\epsilon, \mathcal{F}_{nj}, \|\cdot\|_1)} \sqrt{\Pi(\mathcal{F}_{nj})} \rightarrow 0, \quad (\text{A.1})$$

then $\Pi(f : \|f - f_0\| > \epsilon \mid \mathbf{y}_1, \dots, \mathbf{y}_n) \rightarrow 0$ in \mathbb{P}_0 -probability.

Theorem 3 in [Kruijer et al. \(2010\)](#)

To compute the posterior rate of convergence of the RGM model, we rely on the conditions of Theorem 3 in [Kruijer et al. \(2010\)](#).

Theorem A.2. *Given a statistical model \mathcal{F} with a prior Π , let $(\mathbf{y}_i)_{i=1}^n$ be an i.i.d. sequence with density $f_0 \in \mathcal{F}$. Assume that there exists a sequence of submodels $(\mathcal{F}_n)_{n=1}^{\infty}$ with partitions $\mathcal{F}_n = \bigcup_{j=1}^{\infty} \mathcal{F}_{nj}$, and two sequences $(\underline{\epsilon}_n)_{n=1}^{\infty}, (\bar{\epsilon}_n)_{n=1}^{\infty}$ with $\underline{\epsilon}_n, \bar{\epsilon}_n \rightarrow 0$, $n\underline{\epsilon}_n^2, n\bar{\epsilon}_n^2 \rightarrow \infty$, $\bar{\epsilon}_n \geq \underline{\epsilon}_n$, such that*

$$\Pi(\mathcal{F}_n^c) \lesssim \exp(-4n\underline{\epsilon}_n^2), \quad (\text{A.2})$$

$$\exp(-n\bar{\epsilon}_n^2) \sum_{j=1}^{\infty} \sqrt{\mathcal{N}(\bar{\epsilon}_n, \mathcal{F}_{nj}, \|\cdot\|_1)} \sqrt{\Pi(\mathcal{F}_{nj})} \rightarrow 0, \quad (\text{A.3})$$

$$\Pi\left(f : \int f_0 \log \frac{f_0}{f} \leq \underline{\epsilon}_n^2, \int f_0 \left(\log \frac{f_0}{f}\right)^2 \leq \underline{\epsilon}_n^2\right) \geq \exp(-n\underline{\epsilon}_n^2). \quad (\text{A.4})$$

Then $\Pi(f : \|f - f_0\| > \bar{\epsilon}_n \mid \mathbf{y}_1, \dots, \mathbf{y}_n) \rightarrow 0$ in \mathbb{P}_0 -probability.

B Proof of Theorem 1

Proof. First of all, notice that $h_K(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K) \leq 1$, we see immediately that

$$Z_K \leq \int_{\mathbb{R}^p} \cdots \int_{\mathbb{R}^p} \prod_{k=1}^K p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_k) d\boldsymbol{\mu}_1 \cdots d\boldsymbol{\mu}_K = 1,$$

and hence $-\log Z_K \geq 0$. Now we consider the upper bound for $-\log Z_K$. Suppose h_K is of the form (2.4). Let $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K \stackrel{\text{i.i.d.}}{\sim} p_{\boldsymbol{\mu}}$. Then by Jensen's inequality,

$$-\log Z_K = -\log \mathbb{E} \left[\min_{1 \leq k < k' \leq K} g(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|) \right] \leq \mathbb{E} \left[\max_{1 \leq k < k' \leq K} -\log g(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|) \right].$$

Observing that

$$\left[\max_{1 \leq k < k' \leq K} -\log g(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|) \right]^2 = \max_{1 \leq k < k' \leq K} [\log g(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|)]^2,$$

we obtain

$$\begin{aligned} -\log Z_K &\leq \left\{ \mathbb{E} \left[\max_{1 \leq k < k' \leq K} [\log g(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|)]^2 \right] \right\}^{\frac{1}{2}} \leq \left\{ \sum_{1 \leq k < k' \leq K} \mathbb{E} [\log g(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|)]^2 \right\}^{\frac{1}{2}} \\ &= \left\{ \frac{1}{2} K(K-1) \mathbb{E} [\log g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)]^2 \right\}^{\frac{1}{2}} \leq c_1 K, \end{aligned}$$

where the constant c_1 can be taken as

$$c_1^2 = \frac{1}{2} \mathbb{E} [\log g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)]^2 = \frac{1}{2} \iint_{\mathbb{R}^p \times \mathbb{R}^p} [\log g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)]^2 p(\boldsymbol{\mu}_1) p(\boldsymbol{\mu}_2) d\boldsymbol{\mu}_1 d\boldsymbol{\mu}_2 < \infty.$$

Now we consider the case where h_K is of the form (2.5). Still let $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K \stackrel{\text{i.i.d.}}{\sim} p(\boldsymbol{\mu})$. Jensen's inequality yields

$$\begin{aligned} -\log Z_K &= -\log \mathbb{E} \left[\prod_{1 \leq k < k' \leq K} g(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|)^{\frac{1}{K}} \right] \leq \sum_{1 \leq k < k' \leq K} \frac{1}{K} \mathbb{E} [-\log g(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|)] \\ &\leq \frac{K-1}{2} \left\{ \mathbb{E} [\log g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)]^2 \right\}^{\frac{1}{2}} \leq c_2 K \end{aligned}$$

for some constant $c_2 > 0$. □

C Proofs of Posterior Consistency

Proof of Lemma 1

Proof. Without loss of generality we assume that \mathcal{T}_1 is non-empty. Clearly, $\mathcal{T}_m \uparrow \mathbb{R}^p \times \mathcal{S}$ and $c_m \downarrow 1$ as $m \rightarrow \infty$ by the monotone continuity of F_0 . Furthermore, $\phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq (2\pi\sigma^2)^{-\frac{p}{2}}$. Hence, $f_{F_m} = c_m [\phi_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}) \mathbb{I}_{\mathcal{T}_m}(\boldsymbol{\mu})] * F_0 \rightarrow \phi_{\boldsymbol{\Sigma}} * F_0 = f_0$ by the bounded convergence theorem, implying that $\log \frac{f_0}{f_{F_m}} \rightarrow 0$ as $m \rightarrow \infty$. In order to show $\int f_0 \log \frac{f_0}{f_{F_m}} \rightarrow 0$ as $m \rightarrow \infty$, it suffices to find a dominating function $g(\mathbf{y})$ such that $\left| \log \frac{f_0}{f_{F_m}} \right| \leq g$ for all $m \in \mathbb{N}_+$, and the conclusion is guaranteed by the dominating convergence theorem.

First of all, notice that for all $m \in \mathbb{N}_+$, we have $f_{F_m} \leq c_m \phi_{\boldsymbol{\Sigma}} * F_0 \leq c_1 (2\pi\sigma^2)^{-\frac{p}{2}}$, and thus $f_0 \leq c_1 (2\pi\sigma^2)^{-\frac{p}{2}}$ by letting $m \rightarrow \infty$. It follows that $\log \frac{f_0}{f_{F_m}} \geq \log \frac{f_0}{c_1 (2\pi\sigma^2)^{-\frac{p}{2}}}$. Next, we see that

$$\begin{aligned} f_{F_m}(\mathbf{y}) &= c_m \int_{\mathcal{T}_m} \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) dF_0(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\geq \int_{\mathcal{T}_1} \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) dF_0(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\geq (2\pi\sigma^2)^{-\frac{p}{2}} \int_{\mathcal{T}_1} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \boldsymbol{\mu}\|^2\right) dF_0(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \end{aligned}$$

If $\|\mathbf{y}\| \leq 1$, then $\|\mathbf{y} - \boldsymbol{\mu}\| \leq 2$ as $\|\boldsymbol{\mu}\| \leq 1$, and hence $\exp\left(-\frac{\|\mathbf{y} - \boldsymbol{\mu}\|^2}{2\sigma^2}\right) \geq \exp\left(-\frac{2}{\sigma^2}\right)$; If $\|\mathbf{y}\| > 1$, then $\|\mathbf{y} - \boldsymbol{\mu}\| \leq 2\|\mathbf{y}\|$ as $\|\boldsymbol{\mu}\| \leq \|\mathbf{y}\|$, and hence $\exp\left(-\frac{\|\mathbf{y} - \boldsymbol{\mu}\|^2}{2\sigma^2}\right) \geq \exp\left(-\frac{2\|\mathbf{y}\|^2}{\sigma^2}\right)$.

It follows that

$$f_{F_m}(\mathbf{y}) \geq \xi(\mathbf{y}) := (2\pi\sigma^2)^{-\frac{p}{2}} \begin{cases} \exp\left(-\frac{2}{\sigma^2}\right) F_0(\{\boldsymbol{\mu} : \|\boldsymbol{\mu}\| \leq 1\} \times \mathcal{T}_1), & \text{if } \|\mathbf{y}\| \leq 1, \\ \exp\left(-\frac{2\|\mathbf{y}\|^2}{\sigma^2}\right) F_0(\{\boldsymbol{\mu} : \|\boldsymbol{\mu}\| \leq 1\} \times \mathcal{T}_1), & \text{if } \|\mathbf{y}\| > 1. \end{cases} \quad (\text{C.1})$$

and thus, $\log \frac{f_0}{f_{F_m}} \leq \log \frac{f_0}{\xi}$. In particular, $f_0 \geq \xi$ by letting $m \rightarrow \infty$. Together we have

$$\begin{aligned} \log \frac{f_0}{c_1(2\pi\bar{\sigma}^2)^{-\frac{p}{2}}} &\leq \log \frac{f_0}{f_{F_m}} \leq \log \frac{f_0}{\xi} \\ \implies \left| \log \frac{f_0}{f_{F_m}} \right| &\leq g := \max \left\{ \left| \log \frac{f_0}{c_1(2\pi\bar{\sigma}^2)^{-\frac{p}{2}}} \right|, \log \frac{f_0}{\xi} \right\}. \end{aligned}$$

To show that g is f_0 -integrable, it suffices to verify the f_0 -integrability of $\log f_0$ and $\log \xi$.

Notice that $\log c_1 - (\frac{p}{2}) \log(2\pi\bar{\sigma}^2) \geq \log f_0 \geq \log \xi$, implying

$$|\log f_0| \leq |\log c_1| + \frac{p}{2} |\log(2\pi\bar{\sigma}^2)| + |\log \xi|,$$

it is only left to verify the f_0 -integrability of $\log \xi$. When $\|\mathbf{y}\| \leq 1$, $\log \xi$ is constant, and

when $\|\mathbf{y}\| > 1$, we have

$$\begin{aligned} &\int_{\{\|\mathbf{y}\| \geq 1\}} f_0(\mathbf{y}) |\log \xi(\mathbf{y})| d\mathbf{y} \\ &\leq \frac{p}{2} |\log(2\pi\bar{\sigma}^2)| + |\log F_0(\{\boldsymbol{\mu} : \|\boldsymbol{\mu}\| \leq 1\} \times \mathcal{T}_1)| + \frac{2}{\bar{\sigma}^2} \int_{\{\|\mathbf{y}\| \geq 1\}} \|\mathbf{y}\|^2 f_0(\mathbf{y}) d\mathbf{y} \\ &\leq \frac{p}{2} |\log(2\pi\bar{\sigma}^2)| + |\log F_0(\{\boldsymbol{\mu} : \|\boldsymbol{\mu}\| \leq 1\} \times \mathcal{T}_1)| + \frac{2}{\bar{\sigma}^2} \mathbb{E}_0 \|\mathbf{y}\|^2 < \infty, \end{aligned}$$

where the finiteness of $\mathbb{E}_0 \|\mathbf{y}\|^2$ is guaranteed by condition A1 and Fubini's theorem. Hence $\log \xi$ is f_0 -integrable. \square

Proof of Theorem 2

Proof. By Theorem 1 and Lemma 3 in Wu et al. (2008), it suffices to verify conditions C1, C2, C3, and C4. By Lemma 1, for all $\epsilon > 0$, there exists an integer m such that $F_\epsilon = F_m$ satisfies C1. Noticing that $F_\epsilon \in \text{supp}(\Pi^*)$ automatically holds since $\text{supp}(\Pi^*) = \mathcal{M}(\mathbb{R}^p \times \mathcal{S})$, and that \mathcal{S} itself is compact, we can take $D = \mathcal{T}_m$. For any compact $C \subset \mathbb{R}^p$, take large enough a such that $C \subset \{\mathbf{y} : \|\mathbf{y}\| \leq a\}$. In addition, C3 automatically holds, since $C \times D$ is compact in $\mathbb{R}^p \times \mathbb{R}^p \times \mathcal{S}$, and ϕ is strictly positive. It suffices to verify C2 and C4.

To verify C2, it suffices to show that $\log f_{F_m}$ and $\log \inf_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in D} \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are f_0 -

integrable. Notice that

$$(2\pi\sigma^2)^{-\frac{p}{2}} \geq \inf_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in D} \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \geq \zeta_m(\mathbf{y}) := (2\pi\bar{\sigma}^2)^{-\frac{p}{2}} \begin{cases} \exp\left(-\frac{2m^2}{\sigma^2}\right), & \text{if } \|\mathbf{y}\| \leq m, \\ \exp\left(-\frac{2\|\mathbf{y}\|^2}{\sigma^2}\right), & \text{if } \|\mathbf{y}\| > m, \end{cases}$$

since when $\|\mathbf{y}\| \leq m$ we have $(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \leq \sigma^{-2} \|\mathbf{y} - \boldsymbol{\mu}\|^2 \leq 4\sigma^{-2}m^2$, and when $\|\mathbf{y}\| > m$ we have $(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \leq \sigma^{-2} \|\mathbf{y} - \boldsymbol{\mu}\|^2 \leq 4\sigma^{-2}\|\mathbf{y}\|^2$. It follows that $\log \inf_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in D} \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is f_0 -integrable if $\log \zeta_m$ is integrable. When $\|\mathbf{y}\| \leq m$, ζ_m is a constant, and when $\|\mathbf{y}\| > m$,

$$\int_{\{\|\mathbf{y}\| > m\}} f_0(\mathbf{y}) |\log \zeta_m(\mathbf{y})| d\mathbf{y} \leq \frac{p}{2} |\log(2\pi\bar{\sigma}^2)| + \frac{2}{\sigma^2} \mathbb{E}_0 \|\mathbf{y}\|^2 < \infty.$$

Hence $\log \inf_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in D} \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is f_0 -integrable. Using the ξ function constructed in (C.1) in the proof of **Lemma 1**, we see that $c_1(2\pi\sigma^2)^{-\frac{p}{2}} \geq f_{F_m}(\mathbf{y}) \geq \xi(\mathbf{y})$, and it is proved that $\log \xi(\mathbf{y})$ is f_0 -integrable. It follows that $\log f_{F_m}$ is f_0 -integrable.

To verify C4, given compact C with $C \subset \{\mathbf{y} : \|\mathbf{y}\| \leq a\}$ for some large enough $a > 0$, let

$$E = \left\{ \boldsymbol{\mu} : \|\boldsymbol{\mu}\| \leq \max(a, m) + \max \left[1, \sqrt{2\bar{\sigma}^2 \log \left(\frac{8}{(2\pi\sigma^2)^{\frac{p}{2}} c\epsilon} \right)} \right] \right\} \times \mathcal{S}.$$

Then E contains D in its interior, and E is also compact. Therefore the function $(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mapsto \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ on $C \times E$ is uniformly continuous, and hence, as \mathbf{y} varies over C , the class of functions $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in E \mapsto \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) : \mathbf{y} \in C\}$ is also uniformly equicontinuous. Now we show that $\sup\{\phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) : \mathbf{y} \in C, (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in E^c\} < c\epsilon/4$. Since for any $(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \in C \times E^c$, we have

$$\begin{aligned} & \|\mathbf{y}\| \leq a, \quad \|\boldsymbol{\mu}\| > a + \max \left[1, \sqrt{2\bar{\sigma}^2 \log \left(\frac{8}{(2\pi\sigma^2)^{\frac{p}{2}} c\epsilon} \right)} \right] \\ \implies & \|\mathbf{y} - \boldsymbol{\mu}\| \geq \|\boldsymbol{\mu}\| - \|\mathbf{y}\| \geq \max \left[1, \sqrt{2\bar{\sigma}^2 \log \left(\frac{8}{(2\pi\sigma^2)^{\frac{p}{2}} c\epsilon} \right)} \right], \end{aligned}$$

then we obtain

$$\sup_{(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \in C \times E^c} \phi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq \frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}} \exp \left[-\frac{1}{2\sigma^2} (\|\boldsymbol{\mu}\| - \|\mathbf{y}\|)^2 \right] < \frac{c\epsilon}{4}.$$

The proof is thus completed. \square

Proof of Lemma 2

Proof of Lemma 2. Suppose $\delta > 0$ is given. By Lemma A.4 in Ghosal and Van Der Vaart (2001), there exists an ℓ_1 δ -net \mathcal{I}_0 of Δ^K , such that the cardinality $|\mathcal{I}_0|$ of \mathcal{I}_0 is upper bounded by $(5/\delta)^K$. Now let \mathcal{R}_k be an δ -net of $\{\boldsymbol{\mu}_k : \|\boldsymbol{\mu}_k\|_\infty \in (a_k, b_k]\}$ under the $\|\cdot\|_\infty$ -metric. Clearly, one can make $|\mathcal{R}_k| \leq (b_k/\delta + 1)^p$. Furthermore let \mathcal{S}_{jk} be an δ -net of $\{\sqrt{\lambda_j(\boldsymbol{\Sigma}_k)} : \lambda_j(\boldsymbol{\Sigma}_k) \in [\underline{\sigma}^2, \bar{\sigma}^2]\}$ with cardinality $|\mathcal{S}_{jk}| \leq (\bar{\sigma} - \underline{\sigma})/\delta + 1$ under the $\|\cdot\|_\infty$ -metric. It follows that for all $f_F \in \mathcal{F}_K \left(\prod_{k=1}^K (a_k, b_k] \right)$ with $F = \sum_{k=1}^K w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$, there exists some $\mathbf{w}^* = (w_1^*, \dots, w_K^*) \in \mathcal{I}_0$, $\boldsymbol{\mu}_k^* \in \mathcal{R}_k$, $\lambda_{jk}^* \in \mathcal{S}_{jk}$ for $j = 1, \dots, p$ with $\boldsymbol{\Sigma}_k^* = \mathbf{U} \text{diag}(\lambda_{1k}^*, \dots, \lambda_{pk}^*) \mathbf{U}^T$ for $k = 1, \dots, K$, such that $\sum_{k=1}^K |w_k - w_k^*| < \delta$, $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\| < \sqrt{p} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\|_\infty < \sqrt{p}\delta$, and $|\sqrt{\lambda_j(\boldsymbol{\Sigma}_k)} - \sqrt{\lambda_{jk}^*}| < \delta$ for $j = 1, \dots, p$. Denote $H(f, g)$ to be the Hellinger distance between densities f and g , defined by $H(f, g) = \left(\frac{1}{2} \int (\sqrt{f} - \sqrt{g})^2 \right)^{\frac{1}{2}}$. Observe that

$$\begin{aligned} H(\phi_{\boldsymbol{\Sigma}_k}(\mathbf{y} - \boldsymbol{\mu}_k), \phi_{\boldsymbol{\Sigma}_k^*}(\mathbf{y} - \boldsymbol{\mu}_k^*))^2 &\leq 1 - \prod_{j=1}^p \left(1 - \frac{(\sqrt{\lambda_j(\boldsymbol{\Sigma}_k)} - \sqrt{\lambda_{jk}^*})^2}{\lambda_j(\boldsymbol{\Sigma}_k) + \lambda_{jk}^*} \right)^{\frac{1}{2}} \exp \left(-\frac{\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\|^2}{8\sigma^2} \right) \\ &\leq 1 - \left(1 - \frac{\delta^2}{2\sigma^2} \right)^{\frac{p}{2}} \exp \left(-\frac{p\delta^2}{8\sigma^2} \right) \\ &\leq 1 - \left(1 - \frac{p\delta^2}{2\sigma^2} \right)^{\frac{p}{2}+1} \end{aligned} \tag{C.2}$$

where we use the fact $\exp(-x) \geq 1 - x$ in the last inequality. Denote $F^\star = \sum_{k=1}^K w_k^\star \delta_{(\boldsymbol{\mu}_k^\star, \boldsymbol{\Sigma}_k^\star)}$.

It follows by the triangle inequality that

$$\begin{aligned} \|f_F - f_{F^\star}\|_1 &\leq \sum_{k=1}^K w_k \|\phi_{\boldsymbol{\Sigma}_k}(\mathbf{y} - \boldsymbol{\mu}_k) - \phi_{\boldsymbol{\Sigma}_k^\star}(\mathbf{y} - \boldsymbol{\mu}_k^\star)\|_1 + \sum_{k=1}^K |w_k - w_k^\star| \\ &\leq \sum_{k=1}^K 2\sqrt{2}w_k H(\phi_{\boldsymbol{\Sigma}_k}(\mathbf{y} - \boldsymbol{\mu}_k), \phi_{\boldsymbol{\Sigma}_k^\star}(\mathbf{y} - \boldsymbol{\mu}_k^\star)) + \delta \\ &\leq \delta + 2\sqrt{2} \left[1 - \left(1 - \frac{p\delta^2}{2\sigma^2} \right)^{\frac{p}{2}+1} \right]^{\frac{1}{2}}. \end{aligned}$$

Observing that $\lim_{t \downarrow 0} \frac{1-(1-t)^a}{at} = 1$ holds for $a > 1$, we see that for sufficiently small δ ,

$\|f_F - f_{F^\star}\|_1 \leq C_3\delta$ for some constant $C_3 > 0$, and therefore

$$\begin{aligned} \mathcal{N}\left(C_3\delta, \mathcal{F}_K\left(\prod_{k=1}^K (a_k, b_k]\right), \|\cdot\|_1\right) &\leq \left(\frac{5}{\delta}\right)^K \left(\frac{2(\bar{\sigma} - \underline{\sigma})}{\delta}\right)^{Kp} \prod_{k=1}^K \left(\frac{b_k}{\delta} + 1\right)^p \\ &\leq \frac{\tilde{c}_3^K}{\delta^{Kp+K}} \prod_{k=1}^K \left(\frac{b_k + \delta}{\delta}\right)^p. \end{aligned}$$

for some constant $\tilde{c}_3 > 0$. This yields that

$$\mathcal{N}\left(\delta, \mathcal{F}_K\left(\prod_{k=1}^K (a_k, b_k]\right), \|\cdot\|_1\right) \leq \left(\frac{c_3}{\delta^{2p+1}}\right)^K \left(\prod_{k=1}^K b_k\right)^p$$

for some constant $c_3 > 0$. □

Proof of Lemma 3

Proof. First we need to bound $\sqrt{\Pi(\mathcal{G}_K(\mathbf{a}_K))}$. Recall that $e^{-c_1 K} \leq Z_K \leq 1$ for some constant $c_1 > 0$ by **Theorem 1** and condition A2. We estimate

$$\begin{aligned} \Pi(\mathcal{G}_K(\mathbf{a}_K)) &\leq \Pi(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K : \|\boldsymbol{\mu}_k\| \geq \sqrt{p}a_k, k = 1, \dots, K \mid K) p(K) \\ &\leq \frac{p(K)}{Z_K} \int \dots \int \prod_{k=1}^K \mathbb{I}(\|\boldsymbol{\mu}_k\|^2 \geq pa_k^2) p(\boldsymbol{\mu}_1) d\boldsymbol{\mu}_1 \dots p(\boldsymbol{\mu}_K) d\boldsymbol{\mu}_K \\ &\leq e^{c_1 K} B_2^K \prod_{k=1}^K \exp(-pb_2 a_k^2). \quad (\text{by condition B2 and Theorem 1}) \end{aligned}$$

Now by **Lemma 2** for some constant $c_3 > 0$, we have

$$\mathcal{N}(\delta, \mathcal{G}_K(\mathbf{a}_K), \|\cdot\|_1) \leq \left(\frac{c_3}{\delta^{2p+1}}\right)^K \prod_{k=1}^K (a_k + 1)^p.$$

Hence, by defining $S = \sum_{a_k=0}^{\infty} (a_k + 1)^{\frac{p}{2}} \exp\left(-\frac{pb_2 a_k^2}{2}\right) < \infty$ we estimate

$$\begin{aligned} & \sum_{K=1}^{K_n} \sum_{a_1=0}^{\infty} \cdots \sum_{a_K=0}^{\infty} \sqrt{\mathcal{N}(\delta, \mathcal{G}_K(\mathbf{a}_K), \|\cdot\|_1)} \sqrt{\Pi(\mathcal{G}_K(\mathbf{a}_K))} \\ & \leq \sum_{K=1}^{K_n} \left[\frac{\sqrt{B_2 c_3 e^{c_1}}}{\delta^{p+\frac{1}{2}}} \right]^K \left[\prod_{k=1}^K \sum_{a_k=0}^{\infty} (a_k + 1)^{\frac{p}{2}} \exp\left(-\frac{b_2 p a_k^2}{2}\right) \right] \\ & = \sum_{K=1}^{K_n} \left[\frac{S \sqrt{B_2 c_3 e^{c_1}}}{\delta^{p+\frac{1}{2}}} \right]^K \\ & \leq K_n \left(\frac{M}{\delta^{p+\frac{1}{2}}} \right)^{K_n}, \end{aligned}$$

for some constant $M > 0$ for sufficiently small δ . □

Proof of Theorem 3

Proof. It is sufficient to verify (3.1) and that $\Pi(\mathcal{F}_{K_n}^c) \lesssim \exp(-bn)$ for some $b > 0$, since the KL-property is satisfied. Now take $K_n = \lfloor n/\log n \rfloor$. Then $K_n \log K_n \geq n - \log \log n / \log n \geq n/2$ for large n , which yields $\Pi(\mathcal{F}_{K_n}^c) \lesssim \exp(-B_4 n/2)$ condition B5. Furthermore by **Lemma 3** we have

$$\begin{aligned} & \sum_{K=1}^{K_n} \sum_{a_1=0}^{\infty} \cdots \sum_{a_K=0}^{\infty} \sqrt{\mathcal{N}(\epsilon, \mathcal{G}_K(\mathbf{a}_K), \|\cdot\|_1)} \sqrt{\Pi(\mathcal{G}_K(\mathbf{a}_K))} \\ & \leq \exp \left[\log K_n + K_n \log M + \left(\frac{2p+1}{2} \right) K_n \left(\log \frac{1}{\epsilon} \right) \right] \\ & \leq \exp \left[(p+1) K_n \left(\log \frac{1}{\epsilon} \right) \right] \end{aligned}$$

for sufficiently small ϵ and sufficiently large n . The proof is completed by observing that $(p+1)K_n \log(1/\epsilon) - (4-\tilde{b})n\epsilon^2 \rightarrow -\infty$ as $n \rightarrow \infty$ for any fixed $\epsilon > 0$ and fixed $\tilde{b} \in (0, 4)$. □

D Proofs for Posterior Contraction Rate

Proof of Proposition 1

Proof. Denote $C = 1/(p+1)$. Then by condition B5 we have

$$\Pi(\mathcal{F}_{K_n}^c) = \Pi(K > K_n) \leq \exp(-B_4 K_n \log K_n) \leq \exp[-B_4 C \log C (\log n)^{2t-1}] \leq \exp(-4n\epsilon_n^2)$$

with $t > t_0 + \frac{1}{2}$ for sufficiently large n . Next, by **Lemma 3**

$$\begin{aligned} & \exp(-n\bar{\epsilon}_n^2) \sum_{K=1}^{K_n} \sum_{a_1=0}^{\infty} \cdots \sum_{a_K=0}^{\infty} \sqrt{\mathcal{N}(\bar{\epsilon}_n, \mathcal{G}_K(a_1, \dots, a_K), \|\cdot\|_1)} \sqrt{\Pi(\mathcal{G}_K(a_1, \dots, a_K))} \\ & \leq \exp \left[-(\log n)^{2t} + (p+1)C(\log n)^{2t-1} \left(\frac{1}{2} \log n - t \log \log n \right) \right] \\ & \leq \exp \left[-\frac{1}{2}(\log n)^{2t} \right]. \end{aligned}$$

The RHS of the last display converges to 0 as $n \rightarrow \infty$. □

Proof of Lemma 4

The proof of **Lemma 4** requires the following auxiliary **Lemmas D.1-D.4** that generalize Lemma 3.4, Lemma 4.1, and Lemma 5.1 in [Ghosal and Van Der Vaart \(2001\)](#). Since the proofs are quite similar to those there, we defer them in Section F.

Lemma D.1. *Let F be a probability distribution compactly supported on a subset of $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu}\|_{\infty} \leq a\}$ with $a \lesssim (\log \frac{1}{\epsilon})^{\frac{1}{2}}$. Then for sufficiently small $\epsilon > 0$, there exists a discrete probability distribution F^* on a subset of $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu}\|_{\infty} \leq a\}$ with at most $N \lesssim (\log \frac{1}{\epsilon})^{2p}$ support points, such that $\|f_F - f_{F^*}\|_{\infty} \lesssim \epsilon$, and $\|f_F - f_{F^*}\|_1 \lesssim \epsilon (\log \frac{1}{\epsilon})^{\frac{p}{2}}$.*

Lemma D.2. *Let F be a probability distribution compactly supported on a subset of $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu}\|_{\infty} \leq a\}$ with $a \lesssim (\log \frac{1}{\epsilon})^{\frac{1}{2}}$. Then for sufficiently small $\epsilon > 0$, there exists a discrete probability distribution F^* on $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu}\|_{\infty} \leq 2a\}$ with at most*

$N \lesssim (\log \frac{1}{\epsilon})^{2p}$ support points that are taken from

$$\left\{ (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \frac{\boldsymbol{\mu}}{2\epsilon} \in \mathbb{Z}^p, \frac{\lambda_j(\boldsymbol{\Sigma})}{2\epsilon} \in \mathbb{N}_+, j = 1, \dots, p \right\},$$

such that $\|f_F - f_{F^*}\|_1 \lesssim \epsilon (\log \frac{1}{\epsilon})^{\frac{p}{2}}$.

Lemma D.3. *If $F(\|\boldsymbol{\mu}\| \leq B) > \frac{1}{2}$ for some constant B and F_0 is such that for all $t \geq 0$, $F_0(\|\boldsymbol{\mu}\| > t) \leq \exp(-b't^2)$ for some $b' > 0$, then for $\epsilon = H(f_{F_0}, f_F)$ sufficiently small,*

$$\int f_0 \left(\log \frac{f_0}{f_F} \right)^2 \lesssim \epsilon^2 \left(\log \frac{1}{\epsilon} \right)^2, \quad \int f_0 \log \frac{f_0}{f_F} \lesssim \epsilon^2 \left(\log \frac{1}{\epsilon} \right).$$

Lemma D.4. *Let $\epsilon > 0$ be sufficiently small, $F^* = \sum_{k=1}^N w_k^* \delta_{(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*)}$ be such that $\|\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_{k'}^*\|_\infty \geq 2\epsilon$, and $|\lambda_j(\boldsymbol{\Sigma}_k^*) - \lambda_j(\boldsymbol{\Sigma}_{k'}^*)| \geq 2\epsilon$ whenever $k \neq k'$, $j = 1, \dots, p$. Define*

$$E_k = \left\{ (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu} - \boldsymbol{\mu}_k^*\|_\infty < \frac{\epsilon}{2}, |\lambda_j(\boldsymbol{\Sigma}) - \lambda_j(\boldsymbol{\Sigma}_k^*)| < \frac{\epsilon}{2}, j = 1, \dots, p \right\}.$$

Then for any probability distribution F on $\mathbb{R}^p \times \mathcal{S}$,

$$\|f_F - f_{F^*}\| \lesssim \epsilon + \sum_{k=1}^N |P_F(E_k) - w_k^*|.$$

Proof of Lemma 4. The proof is similar to those in Theorem 5.1 and Theorem 5.2 in Ghosal and Van Der Vaart (2001). First let F'_0 be the re-normalized restriction of F_0 on $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu}\| \leq a\}$. By Lemma A.3 in Ghosal and Van Der Vaart (2001) we obtain $\|f_0 - f_{F'_0}\|_1 \lesssim \epsilon$. Next find $F^* = \sum_{k=1}^N w_k^* \delta_{(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*)}$ by Lemma D.2 such that $N \lesssim (\log \frac{1}{\epsilon})^{2p}$, $\|f_{F'_0} - f_{F^*}\|_1 \lesssim \epsilon (\log \frac{1}{\epsilon})^{\frac{p}{2}}$,

$$(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*) \in \left\{ (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \frac{\boldsymbol{\mu}}{2\epsilon} \in \mathbb{Z}^p, \frac{\lambda_j(\boldsymbol{\Sigma})}{2\epsilon} \in \mathbb{N}_+, j = 1, \dots, p \right\}, \quad k = 1, \dots, N,$$

and F^* is supported on a subset of $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu}\|_\infty \leq 2a\}$. In addition, we can require that $\int \|\boldsymbol{\mu}\|^2 dF'_0 = \int \|\boldsymbol{\mu}\|^2 dF^*$ and still $N \lesssim (\log \frac{1}{\epsilon})^{2p}$. Now we claim that there exists some constant $\gamma > 0$ such that

$$\left\{ F = \sum_{k=1}^N w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} : (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in E_k, \sum_{k=1}^K |w_k - w_k^*| < \epsilon \right\} \subset \left\{ F : \|f_0 - f_F\|_1 \leq \gamma \epsilon \left(\log \frac{1}{\epsilon} \right)^{\frac{p}{2}} \right\}.$$

Suppose F is in the LHS of the last display. Observing that $F(E_k) = w_k$, by **Lemma D.4**, F must satisfy $\|f_F - f_{F^*}\|_1 \lesssim \epsilon$. By the construction of F^* and F'_0 , $\|f_{F'_0} - f_{F^*}\|_1 \lesssim \epsilon \left(\log \frac{1}{\epsilon}\right)^{\frac{p}{2}}$, and $\|f_{F'_0} - f_0\|_1 \lesssim \epsilon$. The result follows from the triangle inequality.

Now still let F be on the LHS of the last display. Observe that $H(f_0, f_F) \lesssim \|f_F - f_0\|_1^{\frac{1}{2}} \lesssim \epsilon^{\frac{1}{2}} \left(\log \frac{1}{\epsilon}\right)^{\frac{p}{4}}$. Let $B = 2 \left(\int \|\boldsymbol{\mu}\|^2 dF_0\right)^{\frac{1}{2}}$. It follows that

$$F^*(\|\boldsymbol{\mu}\| > B) \leq \frac{1}{B^2} \int \|\boldsymbol{\mu}\|^2 dF^* = \frac{1}{B^2} \int \|\boldsymbol{\mu}\|^2 dF'_0 \leq \frac{1}{B^2} \int \|\boldsymbol{\mu}\|^2 dF_0 = \frac{1}{4},$$

where the second equality is due to the requirement $\int \|\boldsymbol{\mu}\|^2 dF'_0 = \int \|\boldsymbol{\mu}\|^2 dF^*$, and the last inequality is because the second moment of F'_0 is no greater than that of F_0 . Therefore for $\epsilon < \min(B/\sqrt{p}, 1/4)$, we have $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\| \leq \sqrt{p} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\|_\infty < B$, and hence

$$\|\boldsymbol{\mu}_k\| > 2B \implies \|\boldsymbol{\mu}_k^*\| \geq \|\boldsymbol{\mu}_k\| - \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\| > 2B - B = B.$$

Hence

$$\begin{aligned} F(\|\boldsymbol{\mu}\| > 2B) &= \sum_{k=1}^N w_k \mathbb{I}(\|\boldsymbol{\mu}_k\| > 2B) \leq \sum_{k=1}^N |w_k - w_k^*| \mathbb{I}(\|\boldsymbol{\mu}_k\| > 2B) + \sum_{k=1}^N w_k^* \mathbb{I}(\|\boldsymbol{\mu}_k\| > 2B) \\ &< \epsilon + \sum_{k=1}^N w_k^* \mathbb{I}(\|\boldsymbol{\mu}_k\| > 2B) \leq \epsilon + \sum_{k=1}^N w_k^* \mathbb{I}(\|\boldsymbol{\mu}_k^*\| > B) \\ &= \epsilon + F^*(\|\boldsymbol{\mu}_k^*\| > B) \leq \frac{1}{2}. \end{aligned}$$

Hence by **Lemma D.3**, we have

$$\int f_0 \left(\log \frac{f_0}{f_F}\right)^2 \lesssim \epsilon \left(\log \frac{1}{\epsilon}\right)^{\frac{p+4}{2}}, \quad \int f_0 \log \frac{f_0}{f_F} \lesssim \epsilon \left(\log \frac{1}{\epsilon}\right)^{\frac{p+2}{2}} \leq \epsilon \left(\log \frac{1}{\epsilon}\right)^{\frac{p+4}{2}},$$

and, as a consequence,

$$\left\{ f_F : F = \sum_{k=1}^N w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} : (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in E_k, \sum_{k=1}^N |w_k - w_k^*| < \epsilon \right\} \subset B \left(f_0, \eta \epsilon^{\frac{1}{2}} \left(\log \frac{1}{\epsilon}\right)^{\frac{p+4}{4}} \right).$$

□

Proof of Theorem 4

Proof. By **Proposition 1** it suffices to find the prior concentration rate. Motivated by **Lemma 4**, we are interested in finding the prior probability of the following event:

$$\tilde{B}(F^*, \epsilon) := \left\{ f_F : F = \sum_{k=1}^N w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} : (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in E_k, \sum_{k=1}^N |w_k - w_k^*| < \epsilon \right\}.$$

where $F^* = \sum_{k=1}^N w_k^* \delta_{(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*)}$, $\|\boldsymbol{\mu}_k^*\| \leq \kappa \left(\log \frac{1}{\epsilon}\right)^{\frac{1}{2}}$ for $k = 1, \dots, K$ for some $\kappa > 0$, $\|\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_{k'}^*\|_\infty \geq 2\epsilon$, $|\lambda_j(\boldsymbol{\Sigma}_k^*) - \lambda_j(\boldsymbol{\Sigma}_{k'}^*)| \geq 2\epsilon$ whenever $k \neq k'$, $j = 1, \dots, p$, $N \lesssim \left(\log \frac{1}{\epsilon}\right)^{2p}$, and

$$E_k = \left\{ (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu} - \boldsymbol{\mu}_k^*\|_\infty < \frac{\epsilon}{2}, |\lambda_j(\boldsymbol{\Sigma}) - \lambda_j(\boldsymbol{\Sigma}_k^*)| < \frac{\epsilon}{2}, j = 1, \dots, p \right\}.$$

It follows that

$$\Pi(\tilde{B}(F^*, \epsilon)) = \Pi(K = N) \Pi \left(\bigcap_{k=1}^N \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in E_k\} \middle| K = N \right) \Pi(\|\mathbf{w} - \mathbf{w}^*\|_1 < \epsilon \mid K = N),$$

where $\mathbf{w} = (w_1, \dots, w_N)$, $\mathbf{w}^* = (w_1^*, \dots, w_N^*) \in \Delta^N$. Since $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in E_k$ implies $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\| > \epsilon$, for sufficiently small ϵ we see that

$$\bigcap_{k=1}^N \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in E_k\} \subset \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{k=1}^N : h_N(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N) \geq (c_2 \epsilon)^N\}$$

by condition A1 for both $r = 1$ and $r = 2$. Notice that $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\|_\infty < \epsilon/2$ for sufficiently small ϵ implies that

$$\|\boldsymbol{\mu}_k\|_\infty \leq \|\boldsymbol{\mu}_k^*\|_\infty + \frac{\epsilon}{2} \leq 2\kappa \left(\log \frac{1}{\epsilon}\right)^{\frac{1}{2}} \implies \|\boldsymbol{\mu}_k\| \leq 2\kappa \sqrt{p} \left(\log \frac{1}{\epsilon}\right)^{\frac{1}{2}}, \quad (\text{D.1})$$

in which case we have

$$\int_{\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\|_\infty < \epsilon/2} p(\boldsymbol{\mu}_k) d\boldsymbol{\mu}_k \geq B_3 \epsilon^p \exp \left[-b_3 (2\kappa \sqrt{p})^\alpha \left(\log \frac{1}{\epsilon}\right)^{\frac{\alpha}{2}} \right].$$

Hence we may proceed to compute

$$\begin{aligned}
& \Pi \left(\bigcap_{k=1}^N \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in E_k\} \right) \\
& \geq \frac{1}{Z_K} \prod_{k=1}^N \left[\int_{\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\|_\infty < \epsilon/2} c_2 \epsilon p(\boldsymbol{\mu}_k) d\boldsymbol{\mu}_k \right] \prod_{k=1}^N \prod_{j=1}^p \left[\int_{\lambda_j(\boldsymbol{\Sigma}_k^*) - \epsilon/2}^{\lambda_j(\boldsymbol{\Sigma}_k^*) + \epsilon/2} p_\lambda(\lambda_{jk}) d\lambda_{jk} \right] \\
& \geq \prod_{k=1}^N \left\{ c_2 B_3 \epsilon^{p+1} \exp \left[-b_3 (2\kappa\sqrt{p})^\alpha \left(\log \frac{1}{\epsilon} \right)^{\frac{\alpha}{2}} \right] \right\} \left(\epsilon \min_{\sigma^2 \leq \lambda \leq \bar{\sigma}^2} p_\lambda(\lambda) \right)^{Np} \\
& \geq \epsilon^{2Np+N} \left[c_2 B_3 \min_{\sigma^2 \leq \lambda \leq \bar{\sigma}^2} p_\lambda(\lambda)^p \right]^N \exp \left[-b_3 (2\kappa\sqrt{p})^\alpha N \left(\log \frac{1}{\epsilon} \right)^{\frac{\alpha}{2}} \right],
\end{aligned}$$

For sufficiently small $\epsilon > 0$, taking logarithm yields

$$-N \left(\log \frac{1}{\epsilon} \right)^{\frac{\alpha}{2}} \lesssim \log \Pi((\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in E_k, k = 1, \dots, N).$$

Using condition B5 and the fact $N \lesssim (\log \frac{1}{\epsilon})^{2p}$, we may further obtain

$$- \left(\log \frac{1}{\epsilon} \right)^{2p+\frac{\alpha}{2}} \lesssim \log \Pi(K = N) + \log \Pi((\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in E_k, k = 1, \dots, N).$$

By Lemma A.2 in [Ghosal and Van Der Vaart \(2001\)](#), we have

$$- \left(\log \frac{1}{\epsilon} \right)^{2p+1} \lesssim -N \left(\log \frac{1}{\epsilon} \right) \lesssim \log \Pi \left(w_1, \dots, w_N : \sum_{k=1}^N |w_k - w_k^*| < \epsilon \right).$$

Observing that $\alpha \geq 2$, we obtain

$$\exp \left[-c_5 \left(\log \frac{1}{\epsilon} \right)^{2p+\frac{\alpha}{2}} \right] \lesssim \Pi(\tilde{B}(F^*, \epsilon)) \lesssim \Pi \left(B \left(f_0, \eta \epsilon^{\frac{1}{2}} \left(\log \frac{1}{\epsilon} \right)^{\frac{p+4}{4}} \right) \right)$$

for some constant $c_5 > 0$. Since $\log \left[\eta \epsilon^{\frac{1}{2}} \left(\log \frac{1}{\epsilon} \right)^{\frac{p+4}{4}} \right]$ and $\log \epsilon$ are of the same order in the sense that their ratio converges to a positive constant as $\epsilon \rightarrow 0$, we conclude that

$$\exp \left[-c_5 \left(\log \frac{1}{\epsilon} \right)^{2p+\frac{\alpha}{2}} \right] \lesssim \Pi(B(f_0, \epsilon)).$$

Setting $\underline{\epsilon}_n = (\log n)^{t_0}/\sqrt{n}$, $\bar{\epsilon}_n = (\log n)^t/\sqrt{n}$ with $t_0 > p + \frac{\alpha}{4}$, $t > t_0 + \frac{1}{2} > p + \frac{\alpha+2}{4}$, we see that

$$-n\underline{\epsilon}_n^2 = -(\log n)^{2t_0} < -\left(\log \frac{1}{\underline{\epsilon}_n}\right)^{2p+\frac{\alpha}{2}} \lesssim \log \Pi(B(f_0, \underline{\epsilon}_n)).$$

Hence (3.4) is satisfied with $\underline{\epsilon}_n = (\log n)^{t_0}/\sqrt{n}$, $t_0 > p + \frac{\alpha}{4}$. The proof is thus completed by applying Proposition 1 and Theorem 3 in Kruijer et al. (2010). \square

E Proofs for the Model Complexity

Proof of Theorem 5

Proof. Suppose $\mathbf{U} = \mathbf{I}_p$. For convenience we use the following notation (only for the proof of Theorem 5): $\mathbf{y}_{1:n} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$, $\boldsymbol{\mu}_{1:K} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)$, $z_{1:n} = (z_1, \dots, z_n)$, $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})^\top$, $\boldsymbol{\mu}_k = (\mu_{k1}, \dots, \mu_{kp})^\top$, $n_k = \sum_{i=1}^n \mathbb{I}(z_i = k)$, $\mathbf{y}_{(k)j} = (y_{ij} : z_i = k)^\top \in \mathbb{R}^{n_k}$, and $\mathbf{1}_{n_k} = (1, \dots, 1)^\top \in \mathbb{R}^{n_k}$. First we bound the marginal likelihood for $(z_{1:n}, K)$ with $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{k=1}^K$ integrated out:

$$p(\mathbf{y}_{1:n} \mid z_{1:n}, K) = \frac{1}{Z_K} \iint h_K(\boldsymbol{\mu}_{1:K}) \prod_{k=1}^K \left\{ \left[\prod_{i:z_i=k} \phi_{\boldsymbol{\Sigma}_k}(\mathbf{y}_i - \boldsymbol{\mu}_k) \right] p(\boldsymbol{\mu}_k) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}_k) d\boldsymbol{\mu}_k d\boldsymbol{\Sigma}_k \right\}.$$

Denote $\boldsymbol{\Sigma}_k = \text{diag}(\lambda_{k1}, \dots, \lambda_{kp})$. Since $h_K(\boldsymbol{\mu}_{1:K}) \leq 1$ and $Z_K^{-1} \leq e^{c_1 K}$ by Theorem 1, it follows that

$$\begin{aligned} p(\mathbf{y}_{1:n} \mid z_{1:n}, K) &\leq \frac{1}{Z_K} \prod_{k=1}^K \prod_{j=1}^p \left\{ \iint p_{\boldsymbol{\mu}}(\mu_{kj}) p_{\lambda}(\lambda_{kj}) \prod_{i:z_i=k} \phi_{\lambda_{kj}}(y_{ij} - \mu_{kj}) d\mu_{kj} d\lambda_{kj} \right\} \\ &\leq \frac{1}{Z_K} \prod_{k=1}^K \prod_{j=1}^p \left\{ \left(\frac{\bar{\sigma}}{\underline{\sigma}} \right)^{n_k} \int p_{\lambda}(\lambda_{kj}) d\lambda_{kj} \int \phi_{\tau^2}(\mu_{kj}) \prod_{i:z_i=k} \phi_{\bar{\sigma}^2}(y_{ij} - \mu_{kj}) d\mu_{kj} \right\} \\ &\leq e^{c_1 K} \left(\frac{\bar{\sigma}}{\underline{\sigma}} \right)^{np} \prod_{j=1}^p \prod_{k=1}^K \phi(\mathbf{y}_{(k)j} \mid \mathbf{0}, \boldsymbol{\Lambda}_k) \end{aligned}$$

where $\mathbf{\Lambda}_k = \bar{\sigma}^2 \mathbf{I}_{n_k} + \tau^2 \mathbf{1}_{n_k} \mathbf{1}_{n_k}^T$. Notice that $\lambda_{\min}(\mathbf{\Lambda}_k) = \bar{\sigma}^2$ and $\lambda_{\max}(\mathbf{\Lambda}_k) = \bar{\sigma}^2 + \frac{\tau^2}{n_k} \leq \bar{\sigma}^2 + \tau^2$, it follows that

$$\phi(\mathbf{y}_{(k)j} \mid \mathbf{0}, \mathbf{\Lambda}_k) \leq \left(\frac{\bar{\sigma}^2 + \tau^2}{\bar{\sigma}^2} \right)^{\frac{n_k}{2}} \phi(\mathbf{y}_{(k)j} \mid \mathbf{0}, (\bar{\sigma}^2 + \tau^2) \mathbf{I}_{n_k}),$$

and hence

$$\prod_{j=1}^p \prod_{k=1}^K \phi(\mathbf{y}_{(k)j} \mid \mathbf{0}, \mathbf{\Lambda}_k) \leq \left(1 + \frac{\tau^2}{\bar{\sigma}^2} \right)^{\frac{np}{2}} \prod_{i=1}^n \phi(\mathbf{y}_i \mid \mathbf{0}, (\bar{\sigma}^2 + \tau^2) \mathbf{I}_p).$$

Therefore, the marginal likelihood with $z_{1:n}$ integrated out can be bounded by

$$p(\mathbf{y}_{1:n} \mid K) \leq \exp(c_1 K + \tilde{c}n) \prod_{i=1}^n \phi(\mathbf{y}_i \mid \mathbf{0}, (\bar{\sigma}^2 + \tau^2) \mathbf{I}_p)$$

for some constants $\tilde{c} > 0$. Furthermore, by the KL property (**Theorem 2**) and a standard result (for example, Proposition 3.2 in ?), for any $\tilde{\gamma} > 0$, for sufficiently large n , the marginal likelihood $p(\mathbf{y}_{1:n})$ can be lower bounded:

$$p(\mathbf{y}_{1:n}) = \int \prod_{i=1}^n \int \phi(\mathbf{y}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) dF(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Pi(dF) \geq \exp(-n\tilde{\gamma}) \prod_{i=1}^n f_0(\mathbf{y}_i)$$

almost surely with respect to \mathbb{P}_0 . Fix some $\tilde{\gamma} > 0$, together with condition B5, we conclude that for some constant $\gamma > 0$,

$$\begin{aligned} \Pi(K \geq N \mid \mathbf{y}_1, \dots, \mathbf{y}_n) &= \frac{1}{p(\mathbf{y}_{1:n})} \sum_{K=N}^{\infty} p(\mathbf{y}_{1:n} \mid K) p(K) \\ &\leq \exp\left(-\frac{B_4}{2} N \log N + \gamma n\right) \prod_{i=1}^n \frac{\phi(\mathbf{y}_i \mid \mathbf{0}, (\bar{\sigma}^2 + \tau^2) \mathbf{I}_p)}{f_0(\mathbf{y}_i)}. \end{aligned}$$

The proof is thus completed by integrating out $\mathbf{y}_1, \dots, \mathbf{y}_n$ with respect to f_0 .

For the case where \mathbf{U} is any general unitary matrix, the proof goes exactly the same as the above argument with the $\boldsymbol{\mu}, \mathbf{y}$ replaced by $\mathbf{U}^T \boldsymbol{\mu}, \mathbf{U}^T \mathbf{y}$, and notice that $p_{\boldsymbol{\mu}}(\mathbf{U}^T \boldsymbol{\mu}) = p_{\boldsymbol{\mu}}(\boldsymbol{\mu})$. \square

Proof of Corollary 1

Proof. The proof is a simple application of the Markov's inequality. Since $\liminf_{n \rightarrow \infty} K_n/n > 0$, then there exists some $\delta_0 > 0$, such that $K_n \geq \delta_0 n$ for sufficiently large n . Hence,

$$-\frac{B_4}{2} K_n \log K_n + \gamma n \leq \gamma n - \frac{B_4 \delta_0}{2} n (\log \delta_0 n) \rightarrow \infty$$

as $n \rightarrow \infty$. By **Theorem 5**, we have

$$\limsup_{n \rightarrow \infty} \mathbb{E} [\Pi(K \geq K_n \mid \mathbf{y}_1, \dots, \mathbf{y}_n)] \leq \lim_{n \rightarrow \infty} \exp \left(-\frac{B_4}{2} K_n \log K_n + \gamma n \right) = 0.$$

By Markov's inequality, for any $\epsilon > 0$,

$$\mathbb{P}_0 [\Pi(K \geq K_n \mid \mathbf{y}_1, \dots, \mathbf{y}_n) > \epsilon] \leq \frac{\mathbb{E} [\Pi(K \geq K_n \mid \mathbf{y}_1, \dots, \mathbf{y}_n)]}{\epsilon} \rightarrow 0$$

as $n \rightarrow \infty$, and the proof is thus completed. \square

F Proofs of Auxiliary Results for Lemma 4

Proof of Lemma D.1

Proof. The proofs are similar to those in Lemma 3.1, Lemma 3.2, and Lemma 3.4 in [Ghosal and Van Der Vaart \(2001\)](#). Let $M = \max \left\{ 2a, \sqrt{8\bar{\sigma}} \left(\log \frac{1}{\epsilon} \right)^{\frac{1}{2}} \right\}$, and let ϵ be sufficiently small such that $M > 2a$. Then

$$\sup_{\|\mathbf{y}\| \geq M} |f_F(\mathbf{y}) - f_{F^*}(\mathbf{y})| \leq 2\phi_{\Sigma}(M - a) \leq 2\phi_{\Sigma}(M/2) \lesssim \exp(-M^2/(8\bar{\sigma}^2)) = \epsilon,$$

so that it suffices to consider $\|\mathbf{y}\| \leq M$. Denote $Q_{\Sigma}(\mathbf{y}) = \mathbf{y}^T \Sigma^{-1} \mathbf{y}$. By Taylor's expansion we have

$$\left| \phi_{\Sigma}(\mathbf{y} - \boldsymbol{\mu}) - \sum_{j=1}^{J-1} \frac{(-1)^j}{2^j (2\pi)^{\frac{p}{2}}} \det(\Sigma)^{-\frac{1}{2}} Q_{\Sigma}^j(\mathbf{y} - \boldsymbol{\mu}) \right| \lesssim \left(\frac{e/2 Q_{\Sigma}(\mathbf{y} - \boldsymbol{\mu})}{J} \right)^J.$$

Hence for any probability distribution F^\star on $\{\boldsymbol{\mu} : \|\boldsymbol{\mu}\| \leq a\} \times \mathcal{S}$, a standard argument of triangle inequality yields

$$\begin{aligned}
\sup_{\|\mathbf{y}\| \leq M} |f_F(\mathbf{y}) - f_{F^\star}(\mathbf{y})| &\leq \sup_{\|\mathbf{y}\| \leq M} \left| \sum_{j=1}^{J-1} \frac{(-1)^j}{2^j (2\pi)^{\frac{p}{2}}} \int \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} Q_{\boldsymbol{\Sigma}}^j(\mathbf{y} - \boldsymbol{\mu}) (dF - dF^\star) \right| \\
&\quad + 2 \sup_{\|\mathbf{y}\| \leq M, \|\boldsymbol{\mu}\| \leq a} \left| \phi_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}) - \sum_{j=1}^{J-1} \frac{(-1)^j}{2^j (2\pi)^{\frac{p}{2}}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} Q_{\boldsymbol{\Sigma}}^j(\mathbf{y} - \boldsymbol{\mu}) \right| \\
&\leq \sup_{\|\mathbf{y}\| \leq M} \left| \sum_{j=1}^{J-1} \frac{(-1)^j}{2^j (2\pi)^{\frac{p}{2}}} \int \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} Q_{\boldsymbol{\Sigma}}^j(\mathbf{y} - \boldsymbol{\mu}) (dF - dF^\star) \right| \\
&\quad + 2c_1 \sup_{\|\mathbf{y}\| \leq M, \|\boldsymbol{\mu}\| \leq a} \left(\frac{e/2 Q_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu})}{J} \right)^J, \tag{F.1}
\end{aligned}$$

for some constant $c_1 > 0$. Suppose $\mathbf{U} = \mathbf{I}_p$. Expanding $Q_{\boldsymbol{\Sigma}}^j(\mathbf{y} - \boldsymbol{\mu})$ by multinomial theorem:

$$Q_{\boldsymbol{\Sigma}}^j(\mathbf{y} - \boldsymbol{\mu}) = \sum_{\substack{r+s+t=j \\ r_1+\dots+r_p=r \\ t_1+\dots+t_p=t \\ s_1+\dots+s_p=s}} \left(\binom{j}{r_1 \dots r_p, s_1 \dots s_p, t_1 \dots t_p} \prod_{i=1}^p y_i^{2r_i} \right) \left(\prod_{i=1}^p \frac{\mu_i^{s_i+2t_i}}{\lambda_i^{r_i+s_i+t_i}(\boldsymbol{\Sigma})} \right).$$

In order that the first term on the RHS of (F.1) vanishes, it is sufficient that

$$\int \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} Q_{\boldsymbol{\Sigma}}^j(\mathbf{y} - \boldsymbol{\mu}) (dF - dF^\star) = 0$$

for all $j = 0, 1, \dots, J-1$. By the multinomial expansion, a sufficient condition for the last display is that

$$\int \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \prod_{i=1}^p \frac{\mu_i^{s_i+2t_i}}{\lambda_i^{r_i+s_i+t_i}(\boldsymbol{\Sigma})} (dF' - dF^\star) = 0$$

for all possible $r_i, s_i, t_i, i = 1, \dots, p$. According to Lemma A.1 in Ghosal and Van Der Vaart (2001), F^\star can be select to be a discrete distribution with at most $N \lesssim J^p(2J-1)^p + 1 \lesssim J^{2p}$ support points. For the case \mathbf{U} is not the identity matrix, the above argument can be applied with y_i and μ_i replaced by $(\mathbf{U}^\top \mathbf{y})_i$ and $(\mathbf{U}^\top \boldsymbol{\mu})_i$, respectively.

Now we focus on the selection of J . Notice that

$$\sup_{\|\mathbf{y}\| \leq M, \|\boldsymbol{\mu}\| \leq a} Q_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}) \lesssim \sup_{\|\mathbf{y}\| \leq M, \|\boldsymbol{\mu}\| \leq a} \|\mathbf{y} - \boldsymbol{\mu}\|^2 \lesssim M^2 \lesssim \left(\log \frac{1}{\epsilon} \right).$$

Hence the second term on the RHS of (F.1) is upper bounded by a constant multiple of $((c_2 \log \frac{1}{\epsilon})/J)^J$ for some constant $c_2 > 0$. Set $J = \lceil (1 + c_2) (\log \frac{1}{\epsilon}) \rceil$. Then

$$\sup_{\|\mathbf{y}\| \leq M} |f_F(\mathbf{y}) - f_{F^*}(\mathbf{y})| \lesssim \left(\frac{(c_2 \log \frac{1}{\epsilon})}{J} \right)^J \lesssim \left(\frac{c_2}{1 + c_2} \right)^{(1+c_2) \log(1/\epsilon)} = \epsilon^{(1+c) \log(1+1/c)} \leq \epsilon$$

for sufficiently small $\epsilon > 0$, where the last inequality is due to the fact $(1 + c) \log(1 + 1/c)$ decrease with c and converges to 1 as $c \rightarrow \infty$. Hence the number N of support points for discrete F^* such that $\|f_F - f_{F^*}\|_\infty \lesssim \epsilon$ is of order $J^{2p} \propto (\log \frac{1}{\epsilon})^{2p}$.

For the inequality regarding L_1 distance, notice that for $\|\mathbf{y}\| > T \geq 2a$, $f_F(\mathbf{y}) \lesssim \exp(-\|\mathbf{y}\|^2/8\bar{\sigma}^2)$, so that

$$\begin{aligned} \|f_F - f_{F^*}\|_1 &\lesssim \int_{\|\mathbf{y}\| > T} \exp\left(-\frac{\|\mathbf{y}\|^2}{8\bar{\sigma}^2}\right) d\mathbf{y} + \int_{\|\mathbf{y}\| < T} \|f_F - f_{F^*}\|_\infty d\mathbf{y} \\ &\lesssim \exp\left(-\frac{T^2}{8\bar{\sigma}^2}\right) + T^p \|f_F - f_{F^*}\|_\infty. \end{aligned} \quad (\text{F.2})$$

Now take

$$T = \max \left\{ 2a, \bar{\sigma} \sqrt{8 \log \left(\frac{1}{\|f_F - f_{F^*}\|_\infty} \right)} \right\}.$$

It follows that the first term on the RHS of (F.2) is bounded by $\|f_F - f_{F^*}\|_\infty \lesssim \epsilon$, while the second term is bounded by a multiple of

$$\|f_F - f_{F^*}\|_\infty \max \left\{ a^p, \log \left(\frac{1}{\|f_F - f_{F^*}\|_\infty} \right)^{\frac{p}{2}} \right\} \lesssim \epsilon \left(\log \frac{1}{\epsilon} \right)^{\frac{p}{2}}.$$

Therefore, for sufficiently small $\epsilon > 0$, $\|f_F - f_{F^*}\|_1 \lesssim \epsilon \left(\log \frac{1}{\epsilon} \right)^{\frac{p}{2}}$. \square

Proof of Lemma D.2

Proof. First for a given ϵ , obtain F' by Lemma D.1 with at most $n \lesssim (\log \frac{1}{\epsilon})^{\frac{1}{2}}$ support points. Write $F' = \sum_k w_k \delta_{(\mu_k, \Sigma_k)}$. For each k , find $\mu_k^* \in \{\mu : \mu/(2\epsilon) \in \mathbb{Z}^p\}$, $\Sigma_k^* \in \{\Sigma : \lambda_j(\Sigma)/(2\epsilon) \in \mathbb{N}_+, j = 1, \dots, p\}$ such that $\|\mu_k - \mu_k^*\| \lesssim \epsilon$ and $\|\Sigma_k - \Sigma_k^*\| \lesssim \epsilon$. Furthermore the function class $\{(\mu, \Sigma) \mapsto \phi(\mathbf{y} \mid \mu, \Sigma)\}_{\mathbf{y} \in \mathbb{R}^p}$ indexed by $\mathbf{y} \in \mathbb{R}^p$ is uniformly Lipschitz

continuous, since $\nabla_{\boldsymbol{\mu}} \phi_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu})$ is uniformly bounded and $\boldsymbol{\Sigma} \in \mathcal{S}$ is compact. Therefore, by taking $F^* = \sum_k w_k \delta_{(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*)}$, we have by the triangle inequality

$$\begin{aligned} \|f_F - f_{F^*}\|_{\infty} &\leq \|f_F - f_{F'}\|_{\infty} + \sum_{k=1}^K w_k \|\phi_{\boldsymbol{\Sigma}_k}(\mathbf{y} - \boldsymbol{\mu}_k) - \phi_{\boldsymbol{\Sigma}_k^*}(\mathbf{y} - \boldsymbol{\mu}_k^*)\|_{\infty} \\ &\lesssim \epsilon + \sum_{k=1}^K w_k L (\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\| + \|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k^*\|) \lesssim \epsilon \end{aligned}$$

where L is the (uniform) Lipschitz constant for the function class $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mapsto \phi_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu})\}_{\mathbf{y} \in \mathbb{R}^p}$.

Now applying the exactly same argument used in deriving (F.2) yields $\|f_F - f_{F^*}\|_1 \lesssim \epsilon (\log \frac{1}{\epsilon})^{\frac{p}{2}}$. \square

Proof of Lemma D.3

Proof. Since $f_0(\mathbf{y}) \leq \underline{\sigma}^p \phi_{\mathbf{I}_p}(\mathbf{0})$, and

$$f_F(\mathbf{y}) \geq \frac{1}{\bar{\sigma}^p} \int_{\{\|\boldsymbol{\mu}\| \leq B\}} \phi_{\mathbf{I}_p} \left(\frac{\mathbf{y} - \boldsymbol{\mu}}{\underline{\sigma}} \right) dF \geq \frac{1}{2\bar{\sigma}^p} \phi_{\mathbf{I}_p} \left(\frac{\mathbf{y}(\|\mathbf{y}\| + B)}{\|\mathbf{y}\|\underline{\sigma}} \right),$$

then we see that $f_0/f_F \lesssim \exp(b_1 \|\mathbf{y}\|^2)$ for some constant $b_1 > 0$. Hence for sufficientl small $\delta > 0$,

$$\int \left(\frac{f_0(\mathbf{y})}{f_F(\mathbf{y})} \right)^{\delta} f_0(\mathbf{y}) d\mathbf{y} \lesssim \int \int \exp(\delta b_1 \|\mathbf{y}\|^2) \exp \left(-\frac{1}{2\underline{\sigma}^2} \|\mathbf{y} - \boldsymbol{\mu}\|^2 \right) dF_0 d\mathbf{y} < \infty. \quad (\text{F.3})$$

The proof is completed by applying Theorem 5 in ?. \square

Proof of Lemma D.4

Proof. Let $E_0 = (\bigcup_k E_k)^c$. We estimate

$$\begin{aligned} |f_F(\mathbf{y}) - f_{F^*}(\mathbf{y})| &\leq \int_{E_0} \phi_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}) dF + \sum_{k=1}^N \int_{E_k} |\phi_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}) - \phi_{\boldsymbol{\Sigma}_k^*}(\mathbf{y} - \boldsymbol{\mu}_k^*)| dF \\ &\quad + \sum_{k=1}^N \phi_{\boldsymbol{\Sigma}_k^*}(\mathbf{y} - \boldsymbol{\mu}_k^*) |P_F(E_k) - w_k^*|. \end{aligned} \quad (\text{F.4})$$

For $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in E_k$, we see that $\|\boldsymbol{\mu} - \boldsymbol{\mu}_k^*\| \lesssim \epsilon$ and $|\lambda_j(\boldsymbol{\Sigma}) - \lambda_j(\boldsymbol{\Sigma}_k^*)| \lesssim \epsilon$. Since eigenvalues of covariance matrices are bounded away from 0 and ∞ , we see that $\left| \sqrt{\lambda_j(\boldsymbol{\Sigma})} - \sqrt{\lambda_j(\boldsymbol{\Sigma}_k^*)} \right| \lesssim \epsilon / \left| \sqrt{\lambda_j(\boldsymbol{\Sigma})} + \sqrt{\lambda_j(\boldsymbol{\Sigma}_k^*)} \right| \lesssim \epsilon$. Hence by (C.2) and the relation between Hellinger distance and $\|\cdot\|_1$, we have $\|\phi_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}) - \phi_{\boldsymbol{\Sigma}_k^*}(\mathbf{y} - \boldsymbol{\mu}_k^*)\|_1 \lesssim \epsilon$ whenever $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in E_k$ for all k and all sufficiently small ϵ . Thus we obtain from Fubini's theorem that

$$\begin{aligned} \|f_F - f_{F^*}\|_1 &\leq \int_{E_0} \int \phi_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}) d\mathbf{y} dF + \sum_{k=1}^N \int_{E_k} \|\phi_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}) - \phi_{\boldsymbol{\Sigma}_k^*}(\mathbf{y} - \boldsymbol{\mu}_k^*)\|_1 dF \\ &\quad + \sum_{k=1}^N |F(E_k) - w_k^*| \int \phi_{\boldsymbol{\Sigma}_k^*}(\mathbf{y} - \boldsymbol{\mu}_k^*) d\mathbf{y} \\ &\lesssim \left[\sum_{k=1}^N w_k^* - \sum_{k=1}^N F(E_k) \right] + \epsilon + \sum_{k=1}^N |F(E_k) - w_k^*| \lesssim \epsilon + \sum_{k=1}^N |F(E_k) - w_k^*|. \end{aligned}$$

□

G Derivation of the Generalized Urn Model

As shown in Miller and Harrison (2016), the marginal distribution of \mathcal{C}_n with K and $z = (z_1, \dots, z_n)$ marginalized out is given by

$$p(\mathcal{C}_n) = V_n(|\mathcal{C}_n|) \prod_{c \in \mathcal{C}_n} \frac{\Gamma(\beta + |c|)}{\Gamma(\beta)} \quad (\text{G.1})$$

where

$$V_n(t) := \sum_{K=t}^{\infty} \frac{\Gamma(K+1)\Gamma(\beta K+1)}{\Gamma(K-t+1)\Gamma(\beta K+n+1)} p(K). \quad (\text{G.2})$$

The following generalized Bayes rule is useful: If $p(\mathbf{y} | \boldsymbol{\theta}) = \phi(\mathbf{y} | \boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \Pi$, then

$$\Pi(\boldsymbol{\theta} \in A | \mathbf{y}) = \int_A \phi(\mathbf{y} | \boldsymbol{\theta}) \Pi(d\boldsymbol{\theta}) / \int \phi(\mathbf{y} | \boldsymbol{\theta}) \Pi(d\boldsymbol{\theta}) \propto \int_A \phi(\mathbf{y} | \boldsymbol{\theta}) \Pi(d\boldsymbol{\theta}). \quad (\text{G.3})$$

Proof of Theorem 6. The restaurant process for the exchangeable partition model pro-

posed by [Miller and Harrison \(2016\)](#) is given by

$$\begin{aligned}\Pi(\mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\} \mid \mathcal{C}_{n-1}) &\propto \frac{V_n(\ell+1)}{V_n(\ell)}\beta \\ \Pi(\mathcal{C}_n = (\mathcal{C}_{n-1} \setminus \{c\}) \cup \{c \cup \{n\}\} \mid \mathcal{C}_{n-1}) &\propto |c| + \beta\end{aligned}$$

where $|\mathcal{C}_{n-1}| = \ell$. Then for any measurable A , the following derivation using chain rule of conditional distributions is available

$$\begin{aligned}&\Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}) \\&= \sum_{\mathcal{C}_n} \Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}, \mathcal{C}_n) p(\mathcal{C}_n \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}) \\&= \sum_{\mathcal{C}_n} \Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}, \mathcal{C}_n) p(\mathcal{C}_n \mid \mathcal{C}_{n-1}) \\&\propto \left[\frac{V_n(\ell+1)\beta}{V_n(\ell)} \right] \Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}, \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}) \\&\quad + \sum_{c \in \mathcal{C}_{n-1}} (|c| + \beta) \Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}, \mathcal{C}_n = (\mathcal{C}_{n-1} \setminus \{c\}) \cup \{c \cup \{n\}\})\end{aligned}$$

Since $\Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}, \mathcal{C}_n = (\mathcal{C}_{n-1} \setminus \{c\}) \cup \{c \cup \{n\}\}) = \delta_{(\boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*)}(A)$, we focus on deriving $\Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}, \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\})$. Since

$$\begin{aligned}&\Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}, \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}) \\&= \sum_{K=\ell+1}^{\infty} \Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}, K) p(K \mid \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}) \\&= \sum_{K=\ell+1}^{\infty} \Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*, c \in \mathcal{C}_{n-1}, K) p(K \mid \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}).\end{aligned}$$

Hence

$$\begin{aligned}&\Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}) \\&\propto \left[\frac{V_n(t+1)\beta}{V_n(t)} \right] \sum_{K=\ell+1}^{\infty} p(K \mid \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}) \Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*, c \in \mathcal{C}_{n-1}, K) \\&\quad + \sum_{c \in \mathcal{C}_{n-1}} (|c| + \beta) \delta_{(\boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*)}(A),\end{aligned} \tag{G.4}$$

and hence, by generalized Bayes rule (G.3),

$$\begin{aligned} \Pi(\boldsymbol{\theta}_n \in A \mid \mathbf{y}_n, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}) &\propto \left[\frac{V_n(t+1)\beta}{V_n(t)} \right] \sum_{K=\ell+1}^{\infty} p(K \mid \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}) \times \\ &\quad \int \int_A \phi(\mathbf{y}_n \mid \boldsymbol{\gamma}_n, \boldsymbol{\Gamma}_n) \Pi(d\boldsymbol{\gamma}_n d\boldsymbol{\Gamma}_n \mid \boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*, c \in \mathcal{C}_{n-1}, K) \\ &\quad + \sum_{c \in \mathcal{C}_{n-1}} (|c| + \beta) \delta_{(\boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*)}(A) \phi(\mathbf{y}_n \mid \boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*). \end{aligned}$$

By definition, for any measurable $A \subset \mathbb{R}^p \times \mathcal{S}$, when $K \geq \ell + 1$, we have

$$\begin{aligned} &\Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*, c \in \mathcal{C}_{n-1}, K) \\ &\propto \iint_A \left[\int \cdots \int h_K(\boldsymbol{\gamma}_c^* : c \in \mathcal{C}_{n-1} \cup \mathcal{C}_{\emptyset}) \prod_{c \in \mathcal{C}_{\emptyset} \setminus \{\underline{c}\}} p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_c^*) d\boldsymbol{\gamma}_c^* \right] p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_{\underline{c}}^*) p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_{\underline{c}}^*) d\boldsymbol{\gamma}_{\underline{c}}^* d\boldsymbol{\Gamma}_{\underline{c}}^*. \\ &= \iint_A L_K(\boldsymbol{\gamma}_{\underline{c}}^*) p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_{\underline{c}}^*) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_{\underline{c}}^*) d\boldsymbol{\gamma}_{\underline{c}}^* d\boldsymbol{\Gamma}_{\underline{c}}^*. \end{aligned}$$

Normalizing the above conditional probability distribution yields

$$\Pi(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*, c \in \mathcal{C}_{n-1}, K) = \frac{\iint_A L_K(\boldsymbol{\gamma}_{\underline{c}}^*) p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_{\underline{c}}^*) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_{\underline{c}}^*) d\boldsymbol{\gamma}_{\underline{c}}^* d\boldsymbol{\Gamma}_{\underline{c}}^*}{\iint L_K(\boldsymbol{\gamma}_{\underline{c}}^*) p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_{\underline{c}}^*) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_{\underline{c}}^*) d\boldsymbol{\gamma}_{\underline{c}}^* d\boldsymbol{\Gamma}_{\underline{c}}^*}. \quad (\text{G.5})$$

Hence the generalized Bayes rule (G.3) yields

$$\Pi(\boldsymbol{\theta}_n \in A \mid \mathbf{y}_n, \boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*, c \in \mathcal{C}_{n-1}, K) = \frac{\iint_A \phi(\mathbf{y}_n \mid \boldsymbol{\gamma}_{\underline{c}}^*, \boldsymbol{\Gamma}_{\underline{c}}^*) L_K(\boldsymbol{\gamma}_{\underline{c}}^*) p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_{\underline{c}}^*) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_{\underline{c}}^*) d\boldsymbol{\gamma}_{\underline{c}}^* d\boldsymbol{\Gamma}_{\underline{c}}^*}{\iint \phi(\mathbf{y}_n \mid \boldsymbol{\gamma}_{\underline{c}}^*, \boldsymbol{\Gamma}_{\underline{c}}^*) L_K(\boldsymbol{\gamma}_{\underline{c}}^*) p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_{\underline{c}}^*) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_{\underline{c}}^*) d\boldsymbol{\gamma}_{\underline{c}}^* d\boldsymbol{\Gamma}_{\underline{c}}^*}.$$

Notice that, again, by the generalized Bayes rule (G.3), we have

$$\begin{aligned}
& p(K \mid \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}) \iint_A \phi(\mathbf{y}_n \mid \gamma_n, \Gamma_n) \Pi(d\gamma_n d\Gamma_n \mid \gamma_c^*, \Gamma_c^*, c \in \mathcal{C}_{n-1}, K) \\
&= p(K \mid \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}) \iint \phi(\mathbf{y}_n \mid \gamma_n, \Gamma_n) \Pi(d\gamma_n d\Gamma_n \mid \gamma_c^*, \Gamma_c^*, c \in \mathcal{C}_{n-1}, K) \times \\
&\quad \Pi(\theta_n \in A \mid \mathbf{y}_n, \gamma_c^*, \Gamma_c^*, c \in \mathcal{C}_{n-1}, K) \\
&= p(K \mid \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}) \left[\frac{\iint \phi(\mathbf{y}_n \mid \gamma_c^*, \Gamma_c^*) L_K(\gamma_c^*) p_\mu(\gamma_c^*) p_\Sigma(\Gamma_c^*) d\gamma_c^* d\Gamma_c^*}{\iint L_K(\gamma_c^*) p_\mu(\gamma_c^*) p_\Sigma(\Gamma_c^*) d\gamma_c^* d\Gamma_c^*} \right] \times \\
&\quad \Pi(\theta_n \in A \mid \mathbf{y}_n, \gamma_c^*, \Gamma_c^*, c \in \mathcal{C}_{n-1}, K) \\
&= p(K \mid \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}) \times \\
&\quad \left[\frac{\int \cdots \iint \phi(\mathbf{y}_n \mid \gamma_c^*, \Gamma_c^*) h_K(\gamma_c^* : c \in \mathcal{C}_{n-1} \cup \mathcal{C}_\emptyset) p_\Sigma(\Gamma_c^*) d\Gamma_c^* \prod_{c \in \mathcal{C}_\emptyset} p_\mu(\gamma_c^*) d\gamma_c^*}{\int \cdots \int h_K(\gamma_c^* : c \in \mathcal{C}_{n-1} \cup \mathcal{C}_\emptyset) \prod_{c \in \mathcal{C}_\emptyset} p_\mu(\gamma_c^*) d\gamma_c^*} \right] \times \\
&\quad \left[\frac{\iint_A \phi(\mathbf{y}_n \mid \gamma_c^*, \Gamma_c^*) L_K(\gamma_c^*) p_\mu(\gamma_c^*) d\gamma_c^* d\Gamma_c^*}{\iint \phi(\mathbf{y}_n \mid \gamma_c^*, \Gamma_c^*) L_K(\gamma_c^*) p_\mu(\gamma_c^*) d\gamma_c^* d\Gamma_c^*} \right] \\
&= \alpha_K G_K(A).
\end{aligned}$$

The proof is thus completed. □

H Details of Posterior Inference

In this section we provide the detailed blocked-collapsed Gibbs sampler in **Algorithm 1** when a conjugate prior on the covariance matrices for all components is used: $\Sigma_k = \text{diag}(\lambda_{1k}, \dots, \lambda_{pk})$ and $\lambda_{jk} \stackrel{\text{i.i.d.}}{\sim} p(\lambda) \propto \mathbb{I}(\lambda \in [\bar{\sigma}^{-2}, \underline{\sigma}^{-2}]) \lambda^{-a_0-1} \exp(-b_0/\lambda)$, $j = 1, \dots, p, k = 1, \dots, K$. Easy extension of the sampler is available when one use Inverse-Wishart distribution on the non-diagonal covariance matrices Σ_k 's. An practical issue for the implementation of the Gibbs sampler is sampling $p(K \mid \mathcal{C})$. Using formula (3.7) in [Miller and Harrison \(2016\)](#), we see that $p(K \mid \mathcal{C}) \propto \frac{K!}{(K+n)!(K-|\mathcal{C}|)!}$. Notice that for $K \gg |\mathcal{C}|$, $p(K) \approx 0$, and therefore in practice one may use the following approximate sampling scheme

$$p(K \mid \mathcal{C}) \propto \frac{K!}{(K+n)!(K-|\mathcal{C}|)!}, \quad K = |\mathcal{C}|, |\mathcal{C}| + 1, \dots, |\mathcal{C}| + m$$

for a moderate choice of the perturbation range m , especially when n is large, in which case the probability of having large number of empty components(*i.e.* $K \gg |\mathcal{C}|$) is negligible.

Algorithm 1 Blocked-Collapsed Gibbs Sampler

Input:

- 1: Observations $(\mathbf{y}_i)_{i=1}^n$;
 - 2: Hyperparameters (a_0, b_0) , τ, g_0 , $0 < \underline{\sigma} < \bar{\sigma} < \infty$;
 - 3: Burn-in time B ;
 - 4: Number of posterior samples T ;
 - 5: Guess upper bound K_{\max} on K ;
 - 6: Perturbation range m for approximate sampling $p(K \mid \mathcal{C})$.
 - 7: **Initialize:**
 - 8: Set $\ell = 1$, select $K \leq n$, and sample $\boldsymbol{\mu}_k \sim \text{N}(\mathbf{0}, \tau^2 \mathbf{I}_p)$;
 - 9: For $k = 1, \dots, K$, set $\Sigma_k = \mathbf{I}_p$;
 - 10: For $i = 1, \dots, n$, set $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_k, \Sigma_k)$ if $k = \arg\max_k \phi(\mathbf{y}_i \mid \boldsymbol{\mu}_k, \Sigma_k)$;
 - 11: Compute \mathcal{C} from $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)$;
 - 12: Set $(\boldsymbol{\theta}_1^{(1)}, \dots, \boldsymbol{\theta}_n^{(1)}) = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)$.
-

13: **For** $t_{\text{it}} = 2, \dots, (B + T)$

14: Set $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) = (\boldsymbol{\theta}_1^{(t_{\text{it}}-1)}, \dots, \boldsymbol{\theta}_n^{(t_{\text{it}}-1)})$

15: **For** $i = 1, \dots, n$

16: Set $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) \setminus \{\boldsymbol{\theta}_i\}$, compute \mathcal{C}_{-i} from $\boldsymbol{\theta}_{-i}$, and set $\ell = |\mathcal{C}_{-i}|$.

17: Label $(c : c \in \mathcal{C}_{-i})$ as (c_1, \dots, c_ℓ) ;

18: For $k = 1, \dots, \ell$, set $(\boldsymbol{\gamma}_k^*, \boldsymbol{\Gamma}_k^*) = (\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i)$ if $i \in c_k$.

19: Sample K from

$$p(K | \mathcal{C}) \propto \frac{K!}{(K - \ell)!(K + n)!}, \quad K = \ell, \ell + 1, \dots, \ell + m.$$

20: Sample $\boldsymbol{\Sigma}_K^* = \text{diag}(\lambda_1^*, \dots, \lambda_p^*)$ by $\lambda_j^* \sim \text{Inv-Gamma}(a_0, b_0) \mathbb{I}([\underline{\sigma}^2, \bar{\sigma}^2])$, $j = 1, \dots, p$.

21: For $k = \ell + 1, \dots, K$, sample $\boldsymbol{\gamma}_k^* \sim \text{N}(\mathbf{0}, \tau^2 \mathbf{I}_p)$;

22: Sample $U \sim \text{Unif}(0, 1)$, and compute

$$g(\boldsymbol{\gamma}_1^*, \dots, \boldsymbol{\gamma}_K^*) = \min_{1 \leq k < k' \leq K} \left(\frac{\|\boldsymbol{\gamma}_k^* - \boldsymbol{\gamma}_{k'}^*\|}{g_0 + \|\boldsymbol{\gamma}_k^* - \boldsymbol{\gamma}_{k'}^*\|} \right);$$

23: If $U < g(\boldsymbol{\gamma}_1^*, \dots, \boldsymbol{\gamma}_K^*)$, then accept the new proposed samples;

24: Otherwise go to line NO. 21 and resample.

25: Sample \mathcal{C} according to the categorical distribution

$$\mathbb{P}(\mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\} | -) \propto \left(\frac{V_n(\ell + 1)}{V_n(\ell)} \beta \right) \phi(\mathbf{y}_i | \boldsymbol{\gamma}_K^*, \boldsymbol{\Gamma}_K^*),$$

$$\mathbb{P}(\mathcal{C} = (\mathcal{C}_{-i} \setminus \{c\}) \cup \{c \cup \{i\}\} | -) \propto \phi(\mathbf{y}_i | \boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*) (|c| + \beta), \quad c \in \mathcal{C}_{-i}.$$

26: If $\mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\}$, then set $\boldsymbol{\theta}_i = (\boldsymbol{\gamma}_K^*, \boldsymbol{\Gamma}_K^*)$;

27: If $\mathcal{C} = (\mathcal{C}_{-i} \setminus \{i\}) \cup \{c \cup \{i\}\}$ for some $c \in \mathcal{C}_{-i}$, then set $\boldsymbol{\theta}_i = \boldsymbol{\theta}_c^* = (\boldsymbol{\gamma}_c^*, \boldsymbol{\Gamma}_c^*)$.

28: Change the current state to $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)$ and \mathcal{C} .

29: **End For**

30: Set $\ell = |\mathcal{C}|$, label $(c : c \in \mathcal{C})$ as (c_1, \dots, c_ℓ) .

31: For $k = 1, \dots, \ell$, set $\boldsymbol{\theta}_k^* = (\boldsymbol{\gamma}_k^*, \boldsymbol{\Gamma}_k^*) = (\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i)$ if $i \in c_k$.

32: Sample K from

$$p(K | \mathcal{C}) \propto \frac{K!}{(K - \ell)!(K + n)!}, \quad K = \ell, \ell + 1, \dots, \ell + m.$$

33: For $j = 1, \dots, p$, $k = 1, \dots, \ell$, sample λ_{jk} from

$$(\lambda_{jk}^* | -) \sim \text{Inv-Gamma} \left(a_0 + \frac{|c_k|}{2}, b_0 + \frac{1}{2} \sum_{i \in c_k} (y_{ij} - \gamma_{kj}^*)^2 \right) \mathbb{I}([\underline{\sigma}^2, \bar{\sigma}^2])$$

where $\boldsymbol{\gamma}_k^* = (\gamma_{1k}^*, \dots, \gamma_{pk}^*)^T$.

34: For $k = 1, \dots, K$, set $\boldsymbol{\Gamma}_k^* = \text{diag}(\lambda_{1k}^*, \dots, \lambda_{pk}^*)$.

35: For $k = 1, \dots, K$, sample $\gamma_k^* \sim N(\mathbf{m}_k, \mathbf{V}_k)$ where

$$\mathbf{V}_k = \left(\mathbf{\Gamma}_k^{\star-1} \sum_{i=1}^n \mathbb{I}(\gamma_i = \gamma_k^*) + \frac{2}{\tau^2} \mathbf{I}_p \right)^{-1},$$

$$\mathbf{m}_k = \mathbf{V}_k \left(\mathbf{\Gamma}_k^{\star-1} \sum_{i=1}^n \mathbb{I}(\gamma_i = \gamma_k^*) \mathbf{y}_i \right).$$

36: Sample $U \sim \text{Unif}(0, 1)$ and compute

$$g(\gamma_1^*, \dots, \gamma_K^*) = \min_{1 \leq k < k' \leq K} \left(\frac{\|\gamma_k^* - \gamma_{k'}^*\|}{g_0 + \|\gamma_k^* - \gamma_{k'}^*\|} \right);$$

37: If $U < g(\gamma_1^*, \dots, \gamma_K^*)$, then accept the new proposed samples;

38: Otherwise go to line NO.35 and resample.

39: For $i = 1, \dots, n$, set $\boldsymbol{\theta}_i = (\gamma_i, \mathbf{\Gamma}_i) = (\gamma_k^*, \mathbf{\Gamma}_k^*)$ if $i \in c_k$.

40: Change the current state to $(\boldsymbol{\theta}_1^{(t_{\text{it}})}, \dots, \boldsymbol{\theta}_n^{(t_{\text{it}})}) = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)$.

41: **End For**

Output:

42: Posterior Samples $(\boldsymbol{\theta}_1^{(t_{\text{it}})}, \dots, \boldsymbol{\theta}_n^{(t_{\text{it}})})_{t_{\text{it}}=B+1}^{B+T}$, where $\boldsymbol{\theta}_i^{(t_{\text{it}})} = (\gamma_i^{(t_{\text{it}})}, \mathbf{\Gamma}_i^{(t_{\text{it}})})$, $i = 1, \dots, p$.

I Convergence Diagnostics

Convergence Check for Subsection 5.1

We check convergence via the trace plots and autocorrelations of some randomly selected γ_i 's (which are identifiable compared to the exact means for different components) in Figure 5, showing no signs of non-convergence.

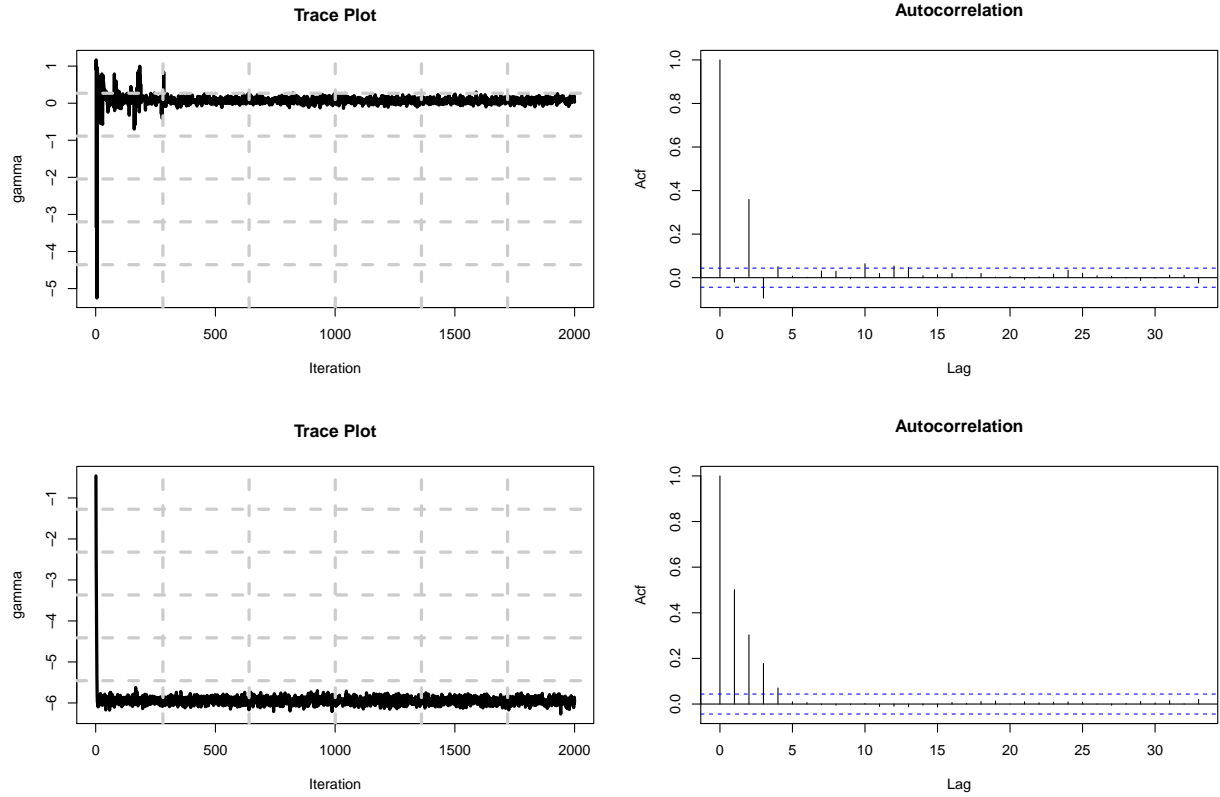


Figure 5: Fitting Multi-Modal Density: The trace plots and the autocorrelation plots of the post-burn-in posterior samples of some randomly selected γ_i 's.

Convergence Check for Subsection 5.2

We check convergence via the trace plots and the autocorrelations of some randomly selected γ_i 's in Figure 5, showing no signs of non-convergence.

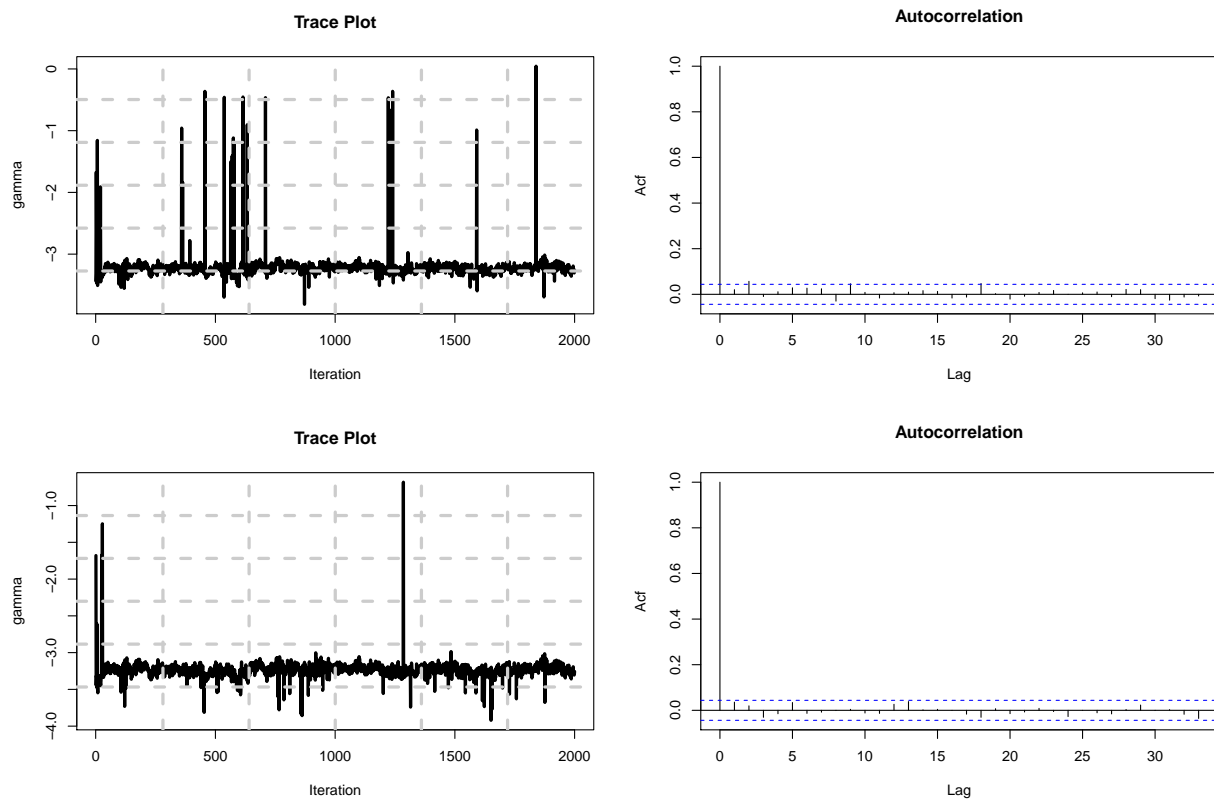


Figure 6: Fitting Uni-Modal Density: The trace plots and the autocorrelation plots of the post-burn-in posterior samples of some randomly selected γ_i 's.

Convergence Check for Subsection 5.2

The trace plots and the autocorrelations of some randomly selected γ_i 's in Figure 7, indicate no signs of non-convergence.

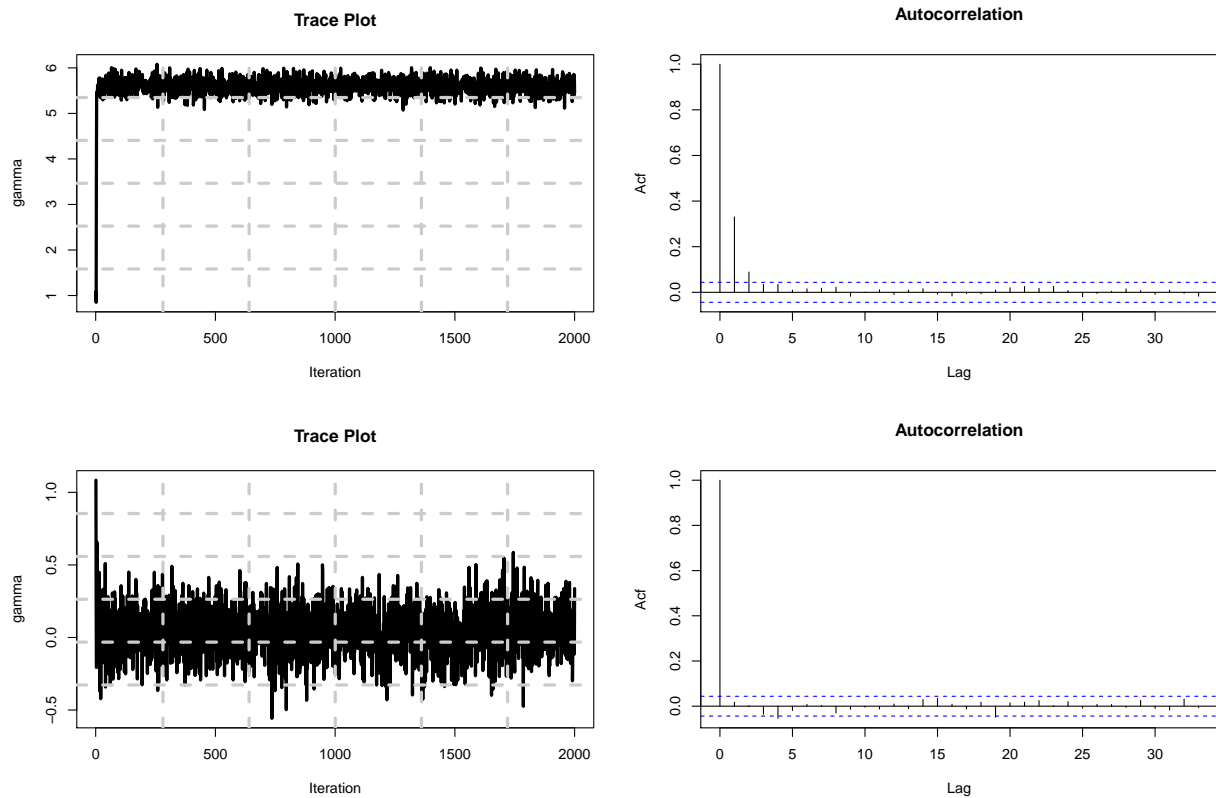


Figure 7: Multivariate Model-Based Clustering: The trace plots and the autocorrelation plots of the post-burn-in posterior samples of some randomly selected γ_i 's.

Convergence Check for Subsection 5.4

The trace plots and the autocorrelations of some randomly selected γ_i 's in Figure 7, indicate no signs of non-convergence.

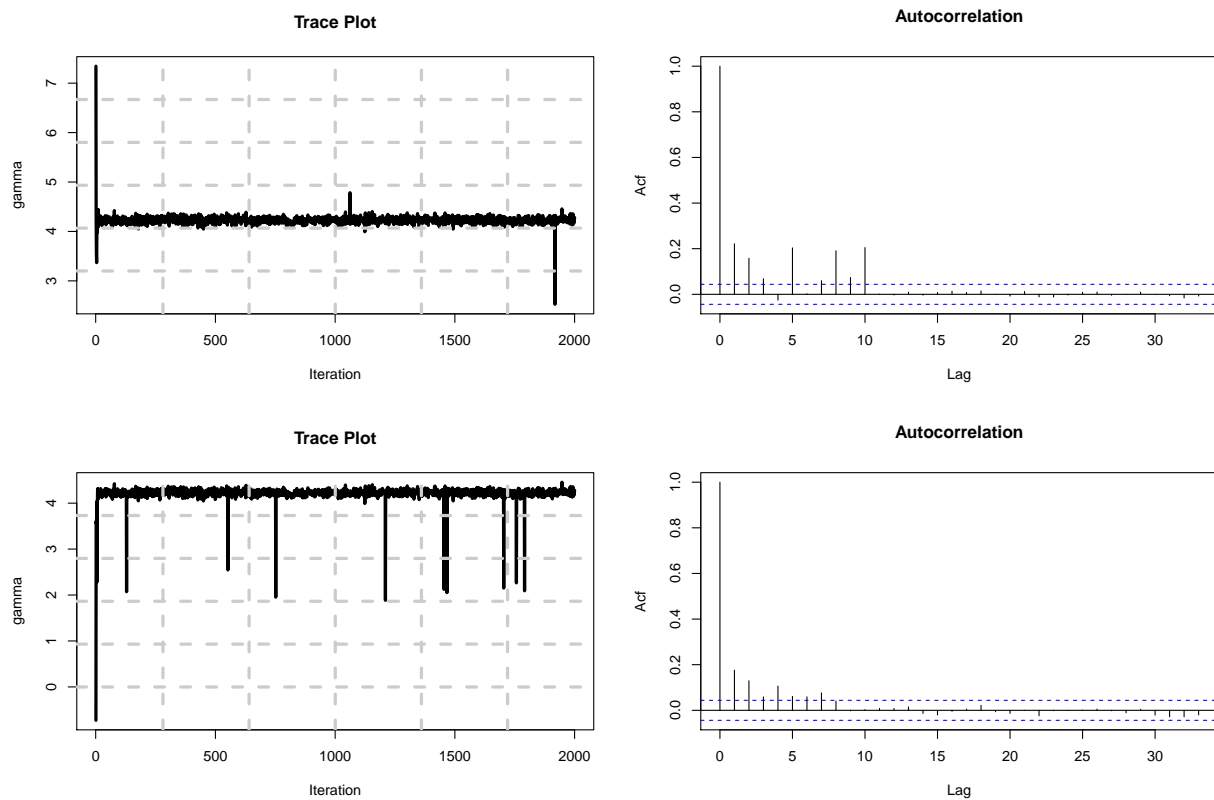


Figure 8: Old Faithful Geyser Eruption Data: Trace plots and autocorrelation plots of the post-burn-in posterior samples of some randomly selected γ_i 's.

References

- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *The Annals of Statistics*, pages 1152–1174.
- Canale, A., De Blasi, P., et al. (2017). Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, 23(1):379–404.
- Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian Inference for Gene Expression and Proteomics*, pages 201–218.
- Fuquene, J., Steel, M., and Rossell, D. (2016). On choosing mixture components via non-local priors. *arXiv preprint arXiv:1604.00314*.
- Garcia-Escudero, L. A. and Gordaliza, A. (1999). Robustness properties of k-means and trimmed k-means. *Journal of the American Statistical Association*, 94(447):956–969.
- Ghosal, S., Ghosh, J. K., Ramamoorthi, R., et al. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158.
- Ghosal, S. and Van Der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29(5):1233–1263.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.
- Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

- Ishwaran, H. and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11(3):508–532.
- Kruijer, W., Rousseau, J., Van Der Vaart, A., et al. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257.
- MacEachern, S. N. and Mueller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.
- Miller, J. W. and Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems*, pages 199–206.
- Miller, J. W. and Harrison, M. T. (2016). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, (just-accepted).
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.
- Nobile, A. (1994). *Bayesian analysis of finite mixture distributions*. PhD thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.
- Petralia, F., Rao, V., and Dunson, D. B. (2012). Repulsive mixtures. In *Advances in Neural Information Processing Systems*, pages 1889–1897.
- Pettit, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 52(1):175–184.
- Qin, Y. and Priebe, C. E. (2013). Maximum Lq-likelihood estimation via the expectation-maximization algorithm: a robust estimation of mixture models. *Journal of the American Statistical Association*, 108(503):914–928.

- Quinlan, J. J., Quintana, F. A., and Page, G. L. (2017). Parsimonious hierarchical modeling using repulsive distributions. *arXiv preprint arXiv:1701.04457*.
- Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26.
- Scricciollo, C. et al. (2011). Posterior rates of convergence for Dirichlet mixtures of exponential power densities. *Electronic Journal of Statistics*, 5:270–308.
- Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.
- Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation*®, 36(1):45–54.
- Walker, S. G., Lijoi, A., Pruenster, I., et al. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics*, 35(2):738–746.
- Wu, Y. and Ghosal, S. (2010). The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 101(10):2411–2419.
- Wu, Y., Ghosal, S., et al. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, 2:298–331.
- Xu, Y., Mueller, P., and Telesca, D. (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics*, 72(3):955–964.